DISSERTATIONS IN
HEALTH
SCIENCES

Jussi Paananen

Bioinformatic Approaches for Integration of Genomic Information

Publications of the University of Eastern Finland
Dissertations in Health Sciences

UNIVERSITY OF
EASTERN FINLAND

JUSSI PAANANEN

Bioinformatic Approaches for Integration of Genomic Information

To be presented by permission of the Faculty of Health Sciences, University of Eastern Finland for public examination in Auditorium ML2, Medistudia building, Kuopio, on Friday, July 6th 2012, at 12 noon

Publications of the University of Eastern Finland
Dissertations in Health Sciences
120

Department of Neurobiology, A.I. Virtanen Institute, Faculty of Health Sciences,
University of Eastern Finland
Kuopio
2012

Kopijyvä Oy
Kuopio, 2012
Finland

Distributor:
University of Eastern Finland
Kuopio Campus Library
P.O.Box 1627
FI-70211 Kuopio, Finland
http://www.uef.fi/kirjasto

ISBN (print): 978-952-61-0836-0
ISBN (pdf): 978-952-61-0837-7
ISSN (print): 1798-5706
ISSN (pdf): 978-952-61-0837-7
ISSN-L: 1798-5706

Author's address: Institute of Clinical Medicine/School of Medicine
University of Eastern Finland
KUOPIO
FINLAND

Supervisors: Professor Garry Wong, Ph.D.
Department of Neurobiology/A.I. Virtanen Institute for Molecular Sciences
University of Eastern Finland
KUOPIO
FINLAND

Markus Storvik, Ph.D.
Pharmacology/School of Pharmacy
University of Eastern Finland
KUOPIO
FINLAND

Reviewers: Professor Mauno Vihinen, Ph.D.
Department of Experimental Medical Science
Lund University
LUND
SWEDEN

Product Manager Eija Korpelainen, Ph.D.
CSC — IT Center for Science
ESPOO
FINLAND

Opponent: Docent Christophe Roos, Ph.D.
Department of Signal Processing
Tampere Technical University
TAMPERE
FINLAND

Paananen, Jussi
Bioinformatic Approaches for Integration of Genomic Information, 43 p.
University of Eastern Finland, Faculty of Health Sciences, 2012
Publications of the University of Eastern Finland. Dissertations in Health Sciences 120. 2012. 43 p.

ISBN (print): 978-952-61-0836-0
ISBN (pdf): 978-952-61-0837-7
ISSN (print): 1798-5706
ISSN (pdf): 978-952-61-0837-7
ISSN-L: 1798-5706

ABSTRACT

Genomic information forms the basis of modern biomedical and biotechnological research. Interpreting this information is a crucial step in order to understand different aspects of life and biology, as well as in developing novel treatments or biotechnological products. One of the most important aspects of interpreting genomic information is the ability to combine and jointly analyze information from various sources, including experiments performed on different species, technologies or levels of genomic information, such as genetics, transcriptomics, proteomics, metabolomics or epigenetics.

To help researchers harness the full potential of genomic information, we have developed novel bioinformatics methods and software tools that can be used to integrate, visualize and analyze genomic data. These include a web-based tool for integrating data for compendium studies, a software tool for 3-D visualization of genomic data, and a human variation database portal. These methods and tools allow researchers to efficiently process large amounts of data from genomic experiments, enabling them to make novel discoveries and hypotheses. In addition, case studies demonstrating these tools are presented.

This thesis also describes and discusses the reasons and challenges of integration of genomic information, while also casting light on the current state of the field together with a review of the existing methods and tools for genomic data integration.

National Library of Medical Classification: QU 26.5, QU 470, QU 58.5, W 26.55.I4
Medical Subject Headings: Bioinformatics; Computational Biology; Database; Data Sharing; Genetics; Genomics; Genomic Structural Variation; Statistical Data Analysis

Paananen Jussi
Bioinformatiikan menetelmiä genomisen tiedon integrointiin, 43 s.
Itä-Suomen yliopisto, terveystieteiden tiedekunta, 2012
Publications of the University of Eastern Finland. Dissertations in Health Sciences 120. 2012. 43 s.

ISBN (print): 978-952-61-0836-0
ISBN (pdf): 978-952-61-0837-7
ISSN (print): 1798-5706
ISSN (pdf): 978-952-61-0837-7
ISSN-L: 1798-5706

TIIVISTELMÄ

Genominen tieto on oleellinen osa nykyaikaista biolääketieteellistä tutkimusta. Tämän tiedon tulkitseminen on tärkeää tutkijoille, jotka pyrkivät ymmärtämään biologiaa ja elämää, samoin kuin tutkijoille, jotka kehittävät uusia hoitomuotoja tai bioteknisiä tuotteita. Yksi tärkeimmistä genomisen tiedon tulkitsemisen keinoista on kyky yhdistää ja analysoida tietoa eri lähteistä. Näitä lähteitä ovat eri lajeilla sekä teknologioilla tehdyt tutkimukset, kuten esimerkiksi geneettiset, transkriptomiset, proteomiset, metabolomiset ja epigeneettiset kokeet.

Auttaaksemme tutkijoita hyödyntämään genomista tietoa, olemme kehittäneet uusia bioinformatiikan menetelmiä ja ohjelmistoja, joita voidaan käyttää genomisen datan yhdistämiseen, visualisointiin ja analysointiin. Nämä uudet menetelmät ja ohjelmistot sisältävät ihmisen geneettistä vaihtelua kuvaavan tietokannan sekä työkaluja genomisen tutkimusdatan integrointiin ja kolmiulotteiseen visualisointiin. Kyseiset menetelmät ja ohjelmistot mahdollistavat laajojen genomisten data-aineistojen käsittelyn ja analysoinnin, auttaen bio- ja lääketieteen tutkijoita tekemään uusia löydöksiä ja hypoteeseja. Teorian ohella väitöskirjassa esitellään kehitettyjä menetelmiä ja ohjelmistoja tapaustutkimusten avulla.

Uusien menetelmien ja ohjelmistojen lisäksi tämä väitöskirja kuvaa ja arvioi genomisen tiedon integrointiin liittyviä syitä ja haasteita, samalla tarkastellen olemassa olevien menetelmien ja työkalujen nykytilaa.

Luokitus: QU 26.5, QU 470, QU 58.5, W 26.55.I4
Yleinen Suomalainen asiasanasto: Bioinformatiikka, genomiikka, integrointi, perinnöllisyystiede, tieto, tiedonlouhinta, tietojenkäsittelytieteet

Acknowledgements

The work presented in this thesis has been carried out in various locations during the years 2004-2012. The work has been sporadic, starting with massive speed, and slowing to almost a complete halt before suddenly being finished to everyone's amazement this year. During this time, an enormous amount of people have helped and contributed to this work, some even positively, and I assume that this is one of the times and places where I should stop and acknowledge you (and at the same time assure you that it won't be the last time or place).

I owe the deepest gratitude to my principal supervisor, Professor Garry Wong, who besides giving me a job back in the day, taught me what science is, and has been a great friend and mentor through all these years. Garry has been a great influence to how I view life, science, and wine.

My other supervisor, Adjunct professor Markus Storvik, deserves my gratitude for not hindering my work with this thesis, and for his insightful comments and feedback (mostly concerning boardgames).

I would like to thank all my co-authors who contributed to the work presented in this thesis. So besides Garry and Markus, thanks go to Robert Ciszek.

During the years this thesis has been in preparation, I have worked at several places. By a quick estimate, during those years, I have had over 200 colleagues working in the same research groups with me. Because I do not want to offend any of them by leaving them out from acknowledgements and because I am sure that I would miss several people who deserve my gratitude, I will rather choose to offend all and everyone of you, and will just simply thank everyone who has worked with me at the research groups of Professor Garry Wong, Academy Professor Markku Laakso, Professor Matti Uusitupa, at the Broad Institute of MIT and Harvard, and at the various other groups that I have collaborated and worked with.

Besides colleagues, my endless gratitude belongs to my friends, who have helped me with various scientific and non-scientific issues during the years. Sometimes the most trivial help, questions or yelling have

helped me to push forward with this thesis (and life in general). Not to mention the non-trivial occasions.

Last and most importantly, I want to thank my family. Thank you for helping me to put things into perspective and to remember how unimportant and silly theses are. And thank you for Everything else as well.

In appreciation of the financial support provided for the work presented in this thesis, I would like to thank the Academy of Finland, Saastamoinen Foundation, Fulbright Program, Instrumentarium Foundation, Olvi Foundation, The Kuopio Naturalists' Society, University of Eastern Finland (UEF), Doctoral Program of Molecular Medicine/UEF, University of Kuopio (UKU), Graduate School of Molecular Medicine/UKU, and Juha Kekäläinen.

Kuopio, June 2012.

- Jussi

List of the original publications

This dissertation is based on the following original publications:

I **Paananen J**, Wong G: Integration of genomic data for pharmacology and toxicology using Internet resources. SAR QSAR Environ Res. 2006 Feb;17(1):25-36.

II **Paananen J**, Storvik M, Wong G: CROPPER: a metagene creator resource for cross-platform and cross-species compendium studies, BMC Bioinformatics. 2006 Sep 22;7:418.

III **Paananen J**, Wong G: FORG3D: force-directed 3D graph editor for visualization of integrated genome scale data, BMC Systems Biology. 2009 Feb 24;3:26.

IV **Paananen J**, Ciszek R, Wong G: Varietas: a functional variation database portal, Database (Oxford). 2010 Jul 29;2010.

Contents

Abbreviations

API	Application programming interface
cDNA	Complementary deoxyribonucleic acid
CGI	Common gateway interface
ChIP	Chromatin immuno-precipitation
ChIP-seq	Chromatin immuno-precipitation sequencing
CNV	Copy number variant
daf-2	Nematode insulin/growth factor receptor gene
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
eQTL	Expression quantitative trait loci
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements coupled with high-throughput sequencing
GEO	Gene Expression Omnibus
GSEA	Gene set enrichment analysis
GUI	Graphical user interface
GWAS	Genome-wide association study
HCA	High-content analysis
HGNC	The HUGO Gene Nomenclature Committee
HGP	Human Genome Project
HUGO	Human Genome Organization
IGV	Integrative Genomics Viewer
IPI	International Protein Index
IRC	Internet Relay Chat
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC-MS	Liquid chromatography-mass spectrometry
LSID	Life Science Identifier
LRG	Locus Reference Genomic
MIAME	Minimum Information About a Microarray Experiment
miRNA	Micro ribonucleic acid

mRNA	Messenger ribonucleic acid
NMR	Nuclear magnetic resonance
NCBI	National Center for Biotechnology Information
pre-mRNA	Precursor messenger ribonucleic acid
RDF	Resource Description Framework
RNA	Ribonucleic acid
RNAi	Ribonucleic acid interference
RNA-seq	Ribonucleic acid sequencing
siRNA	Small interfering ribonucleic acid
SNP	Single nucleotide polymorphism
SOM	Self-organizing map
UCSC	University of California, Santa Cruz
WWW	World Wide Web

1 Introduction

Genomic information forms the basis of modern biotechnology and biomedical research. In their quest for solving the mysteries of nature and life, researchers apply a vast number of different high-throughput research technologies that produce ever-increasing amounts of information. The information itself is of little worth, but the knowledge that can be extracted from it can help us to better understand life and diseases, to create novel therapies and cures, as well as to help us to develop new biotechnology products that improve our everyday lives. Therefore one of the major research challenges of today is the efficient usage of this genomic information. In other words, how to harness the full potential of this information in a cost-efficient way that allows us to find the hidden gems of important knowledge from these mountains of data.

One of the main steps towards the efficient usage of genomic information is integration. Integration of genomic information allows researches to combine and jointly analyze data from various sources and different research technologies, reduce the need to reproduce experiments, and help to reveal knowledge that simply could not be discovered by using information from a single source. To achieve this goal, new bioinformatics methods, software tools and databases are required.

This thesis focuses on the topic of integration of genomic information and is structured in the following way: the second chapter describes different levels of genomic information, the third chapter discusses the reasons why genomic information is integrated and what the related challenges are, the fourth chapter lists the aims of the original publications presented in this thesis, while the fifth and sixth chapters describe the methods and results from those publications. Chapters seven and eight discuss and summarize the thesis.

The original publications presented in this thesis range from a look into existing integration tools and methods (Publication I), to a novel method and software tool for combining data from heterogeneous sources (Publication II), to visualizing integrated genomic information

(Publication III) and to describing a web-based database portal of integrated information for human genomic variation research (Publication IV).

2 Genomic information

Discovery of the structure of DNA and the articulation of the central dogma of molecular biology launched a new era of biological research that culminated in the completion of the Human Genome Project (HGP) (1,2). The post-genomic era that followed has so far yielded enormous amount of new information concerning health, diseases, individual variation and differences between animal and plant species, as well as about numerous other basic concepts of life.

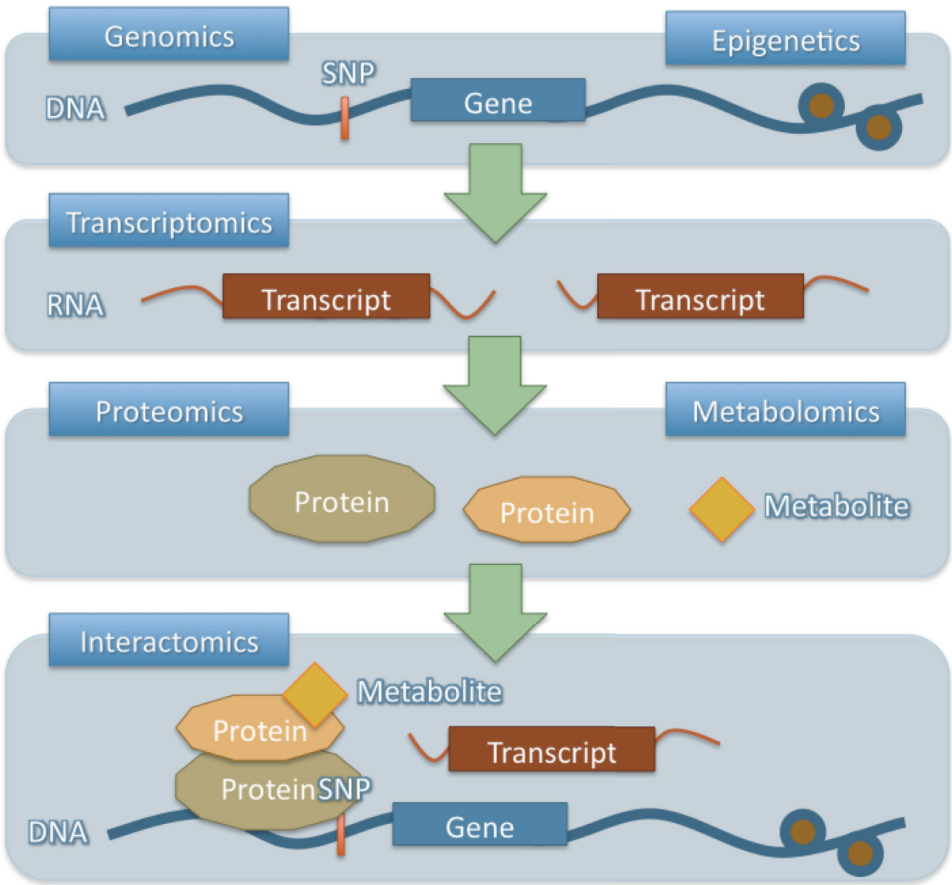

Figure 1. Different levels of biological data and their interactions.

These discoveries in combination with technological advances have opened exciting new possibilities for biological and medical research and

development, but also have created new challenges for extracting knowledge from the constantly growing amount of information. Novel methods and tools are needed to analyze, combine and interpret this information. To accomplish this, understanding of the different levels of underlying biological phenomena and related genomic research technologies is needed. Figure 1 illustrates different levels of genomic information. The subsequent sections will describe these different types of data used in genomic research in greater detail. Table 1 provides an overview of the amount of different types of information in a human genome.

Table 1. Overview of information in a human genome from Ensembl database (Ensembl version 66, *Homo Sapiens* assembly GRCh37.p6).

Type of information	**Amount**
Base pairs	3,286,906,305
Known protein-coding genes	20,563
Novel protein-coding genes	536
Pseudogenes	15,520
RNA genes	11,960
Gene exons	673,807
Gene transcripts	190,053
Short Variants (SNPs, indels, somatic mutations)	52,030,260

2.1 PHENOMICS

Phenotypes are observable characteristics or traits, and therefore represent a classical type of measureable experimental information. Phenotypes can be virtually any type of observable characteristic, including developmental, behavioral, biochemical or physiological properties. Clinical examples of phenotypes include characteristics such as height, eye color or social behavior. Wilhelm Johannsen introduced the terms phenotype and genotype in 1911, making the distinction that phenotypes are observable traits, while genotypes are an organism's

inherited information, even if they are not expressed or directly observable (3). Modern research technologies have enabled direct measurement of genetic information and therefore have made this distinction somewhat ambiguous. In this thesis, phenotype will refer to classical measurable traits, while terms such as transcriptome, metabolome and proteome can be considered as subclasses of phenotypes. The term genotype will denote heritable information such as DNA.

Phenotypes are considered to be results of interactions between genotypes and environment. For example when two tall people have a child, the child is likely to have inherited genotypic qualities that will result in him/her being tall, as long as the environment provides sufficient nutrition. In genomic research, phenotypes are often considered to be the outcomes of interest, e.g. when studying the genetic components that make a person tall, result in high cholesterol levels, cause a disease, or alter behavior. Another common use for phenotypes is to improve analytical models by including phenotypic variables. For example when studying eating habits (a phenotype), one might want to take into account the gender (another phenotype) and weight (yet another phenotype) of the subject. This allows the model to control for differences in eating habits caused by gender or body weight.

As phenotypes can vary from biochemical properties to behavioral habits, the amount of methods for measuring phenotypes is vast, from simple visual inspection and questionnaires to hi-tech research equipment. Phenomics has been defined as the acquisition of high-dimensional phenotypic data on an organism-wide scale (4), essentially meaning studying multiple phenotypes on the level of a single individual.

2.2 GENOMICS AND GENETICS

Genomics, as the study of the genomes of organisms, has strong roots in small scale sequencing studies that have subsequently expanded to full-sized genome projects of different organisms. The undertaking of sequencing and annotating a full genome was previously seen as huge project, requiring large consortia and vast resources. Because of this, the focus has previously been placed on sequencing reference genomes of

individuals, and then using fine-scaled methods to study genomics in large populations. These methods include sequencing of smaller genomic regions, for example regions with candidate genes, and studying of individual variation in the form of single-nucleotide-polymorphism (SNP) genotyping studies.

Common types of current SNP research methods include genome-wide association studies (GWAS) where microarray technologies are used to study up to millions of SNPs identified in projects such as the International HapMap Project (5–7) in large populations. GWAS are usually followed by a replication stage where the most promising findings are replicated in an independent sample (8,9).

Current technological advancements have resulted in creation of novel high-throughput sequencing (so called next-generation sequencing or deep sequencing) technologies. These technologies have the promise of making sequencing large genomic regions, up to full genomes affordable on a population study scale. Thus genomic research is turning back to its roots when more and more focus is shifting back to sequencing studies. A prime example of a population scale whole-genome sequencing projects is the 1000 Genomes project, where the pilot phase included whole-genome sequencing of 179 individuals from four populations, high-coverage sequencing of two mother–father–child trios, and exon-sequencing of 697 individuals from seven populations (10).

2.3 TRANSCRIPTOMICS

The availability of complete genomes has also advanced the study and understanding of the next level of genomic information, the transcriptome. Having the complete genome of an organism allows researchers to gain a better understanding of how genes work and the kind of products created from the genomic sequence through transcription. One of the major advancements based on this information has been the development of high-throughput gene expression analysis technologies, mainly gene expression microarrays, which facilitates measurement of the amount of specific RNA products in a sample on a genome-wide scale. Gene expression microarrays have played a major role in the study of functional genomics, and are currently standard research methods in the field of biotechnology and biomedicine.

Studying the transcriptome has also shed new light on an interesting class of RNAs, the so-called non-coding RNAs that play a role in many cellular processes, but are not necessarily translated into proteins. These include molecules such as microRNAs (miRNAs) and small interfering RNAs (siRNAs) that take part in RNA interference (RNAi), a process that affects regulation of gene expression, and provides the basis for activation and deactivation of genes, and can thefore be used for purposes such as basic research or RNA based therapies (11).

The developments taking place in the field of genomic DNA research are also happening with RNA, as research is moving towards high-throughput sequencing of RNA (RNA-seq) (12). Previous knowledge of the sequence is not required (13), and this provides an improvement over many previous genome-wide transcriptomic research technologies such as gene expression microarrays.

2.4 PROTEOMICS

Similar to transcriptomics, the ability to study the genome sequence of an organism allows the prediction of DNA sequences that are first transcribed into RNA, and later translated into proteins. A large difference between the studying of nucleic acids and proteins is that it has long been understood notion that the three dimensional structure of a protein plays a major role in its function, and knowing the amino acid sequence is just a small part of understanding how these complex structures form. This has also hindered technology developers with additional challenges, as the focus of proteomics research has not been in the sequencing of amino acids, but rather on trying to solve the three-dimensional structure of proteins, and thus how these structures are formed for example by post-translational modifications, and how different proteins or specific regions of proteins interact with other molecules (14).

Besides methods for studying structural aspects of proteins, methods have been developed for studying other areas of proteomics such as expression and localization of proteins (15). These include protein mass spectrometry (16) and protein expression microarrays that work in a fashion similar to DNA and gene expression microarrays, except that

instead of complementary nucleotide sequences, antibodies are used to detect the existence of predefined proteins.

2.5 METABOLOMICS

Metabolomics is a field closely related to proteomics, and shares common research technologies and methods, such as liquid chromatography-mass spectrometry (LC-MS) (17) and nuclear magnetic resonance (NMR) (18) that are used to separate, identify and quantify compounds. Instead of proteins, metabolomics focuses on the study of small-molecule metabolites, such as metabolic intermediates, hormones, and other signaling molecules and secondary metabolites. One of the major focus areas of metabolomics is lipidomics, which aims to comprehensive analyze the lipids present in a biological system (19). Current lipidomics analysis platforms enable quantification of hundreds of different lipid molecules from different lipid classes (20,21).

Besides giving a deeper understanding on how metabolism and biological systems operate, metabolomics as a field of study provides a means for creating diagnostic tools, as well as developing new therapies. Advances have been made for example in the fields of drug metabolism, toxicity, as well as in nutrigenomics (22,23), where the relationship between nutrition and genes is investigated.

2.6 EPIGENETICS

Obtaining the sequence of the human genome also raises the questions concerning mechanisms of inheritance that are not based purely on the DNA sequence of an organism. Epigenetics tries to tackle this question by providing new information on changes that remain through cell division and sometimes across generations, but are not based on DNA sequence (24,25). Mechanisms of epigenetic inheritance include DNA and histone methylation (26), chromatin remodeling (27), as well as inheritance of regulatory molecules such as proteins or RNA involved with RNAi (28). Studying epigenetics has shifted the focus of purely studying the sequence of nucleic acids to also considering the three-dimensional structure (29–31), similar to what has been for a long time central to proteomics studies for a long time.

Central technologies used in studying epigenetics include chromatin immunoprecipitation (ChIP), which is used to study protein DNA interactions. A large-scale microarray variant is ChIP-on-chip. Nonetheless, the field is moving towards high-throughput sequencing technologies such as ChIP-seq, bisulfate sequencing that is used to study DNA methylation and FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements) used to identify open chromatin sites (32).

2.7 INTERACTOMICS

Interactomics describes the study of molecular interactions in cells. Traditionally these studies have been focused on a particular class of interacting molecules, such as protein-protein interactions, but as it is becoming more and more clear, there are important interactions between different levels of genomic molecules, ranging from DNA, RNA to proteins and metabolites. High-throughput studies that combine data from different levels of genomic information are becoming more and more commonplace. Examples of these studies include expression quantitative trait loci (eQTL) analysis where RNA expression is associated with DNA variations (33,34), studies correlating RNA and protein expression (35,36) or studies that combine data from DNA, RNA and metabolite measurements (37).

Systems biology has been one of the main fields to apply interactomics, in order to obtain a comprehensive understanding of a system, such as a cell or an organism. Based on the interactomics information, systems biology aims to build models of the different signaling and metabolic pathways (38).

3 *Integrating genomic information*

In order to gain a more complete understanding of a biological system, integration of different levels of genomic information is required. Moreover, integrating genomic information can be cost-effective and ethical. This chapter discusses the rationale, as well as the related biological and technical challenges for integrating genomic information.

3.1 RATIONALE FOR INFORMATION INTEGRATION

The main scientific rationale for integrating genomic information is simple, the need to relate knowledge between different types of experiments. Living systems consist of numerous levels of interacting components, and to understand the system and related phenomena, one must see what is happening on and between all these different levels. The levels consist of DNA, RNA, proteins, metabolites and other interacting molecules, and because most high-throughput research technologies only focus on a single level (e.g. DNA sequencing or gene expression microarrays) or interactions of two levels (e.g. ChIP-seq measuring protein-DNA binding), researchers need to perform multiple experiments to be able to obtain information from the various levels and their interactions. Combining data from these different experiments is a starting point for getting a deeper understanding of the underlying phenomena of interest.

Often the reasons for information integration are not directly based on the scientific questions, but rather on the practical issues of conducting a study. Examples include performing a gene expression experiment using a specific microarray model, and later desiring to do add more samples to the experiment and noticing that the microarray has been replaced by a newer version. In a situation such as this, it is reasonable to estimate how comparable the results between the different versions of the microarray would be, and how they could be integrated. The same is true when desiring to combine data from experiments performed using similar products from different manufacturers and different technologies

that measure the same biological phenomena (e.g. gene expression microarrays and RNA-seq).

One of the main reasons for wanting to integrate data produced on different technology platforms is to re-use data originally created for other purposes, often by other researchers. Collecting samples and performing high-throughput genomic experiments is expensive, and therefore re-using existing data is often very cost-efficient. As an example, the GeneSapiens project combines data from 9,783 publicly available gene expression samples, representing 175 types of healthy and pathological tissues (39). For a research group to collect and analyze a corresponding number of human tissue samples would mean years of work and an investment of millions of euros.

Re-using and integrating existing data is often also an ethical choice, as it allows studies to be conducted without the need to perform new experiments, reducing the need for collecting new human samples, performing animal experiments or spending research funding. Currently major scientific journals encourage researchers to share their experimental data, and enforce this by requiring researchers to submit their data to public data repositories such as Gene Expression Omnibus (GEO) (40) and ArrayExpress (41) upon publication of the study. Information integration methods also enable researchers to perform studies using model organism or cell lines and project the results to humans (i.e. cross-species studies), making it possible to study phenomena that could not be ethically studied in humans.

3.2 CHALLENGES OF INFORMATION INTEGRATION

Information integration holds many benefits, but has been hampered by several challenges. These challenges can be roughly divided into biological challenges that arise from the complexity of living systems and our limited understanding of them, and technical challenges, which are caused by our technological choices and limitations.

Common biological challenges revolve around differences in study design, experimental conditions and variation between species. For example how comparable are two experiments performed on slightly different time points, or how well does a gene expression experiment performed with mice compare with similar experiment performed with

rats. Biological challenges also include fundamental problems with how we define genomic components such as genes and their relationships to other components. A common example from GWAS studies is a SNP that has been found to be associated with a disease, raising a question about the related pathophysiological process. For this purpose researchers usually look into nearby genes, and often find that the SNP is not located inside any given gene, but might actually be located between two genes. In a situation such as this, if researchers would like to integrate for example gene expression data with the SNP data, there is no clear consensus on which gene(s) the SNP should be linked to.

Table 2. A fraction of identifiers linked to a single human gene (ST7) in the Ensembl database (version 62). Count is the number of identifiers of type named in the Identifier type column.

Identifier type	**Count**	**Examples**
HGNC symbol	1	ST7
Description	1	suppression of tumorigenicity 7
Ensembl Gene ID	1	ENSG00000004866
Illumina HumanWG 6 v3	3	ILMN_1702175, ILMN_1707763, ILMN_1746137
UniProt Gene Name	3	ST7, Q9NRC0, Q9NRC1
RefSeq DNA ID	3	NM_021908, NM_018412, NR_002332
UniProt/TrEMBL Accession	12	Q9NRC0, Q75MZ7, C9JZV9, C9JX79
EMBL (Genbank) ID	18	AC002542, AC106873, AC003987, AF234886
Ensembl Protein ID	21	ENSP00000377092, ENSP00000265437
IPI ID	21	IPI00878915, IPI00878525, IPI00852755, IPI00030166, IPI00922544
Ensembl Transcript ID	29	ENST00000393446, ENST00000265437, ENST00000393451
HGNC transcript name	29	ST7-013, ST7-002, ST7-007, ST7-015, ST7-001
Affy HuEx 1_0 st v2	51	3020553, 3020498, 3020546, 3020497, 3020552
Ensembl Exon ID	82	ENSE00001515400, ENSE00001752339, ENSE00001623906
dbSNP Reference ID	2135	rs72023459, rs71794256, rs58892731, rs71921709

Technological challenges are mainly caused by differences between various genomic research technologies that measure the same experimental variable in different ways, as well as by software tools and databases that identify, structure and manage the experimental data in various ways. For example it is common for a gene to have several dozens of identifiers assigned to it by various consortia, companies and research groups, and the relationships between these identifiers are not often very clear. Table 2 illustrates part of these various identifiers linked to a single gene.

The following paragraphs describe selected challenges in a greater detail.

Amount of data. The amount of data produced by next-generation sequencing has been growing exponentially. A single sequencing machine can now produce over 40 Gb of sequence reads per day, corresponding to over 5 TB of digital image data. The 1000 Genomes project pilot data, representing processed sequences from 629 people is roughly 7.3 TB of sequence data. Downloading this data from a well-connected site is estimated to take from 1-3 weeks. The amount of sequencing data, in addition to increasing the amount of other types of data available from genomic databases is a major challenge from the point of view of data storage, transfer and computational time. (42)

Name space issues. Different genomic databases, research groups and technology providers use different naming methods for describing genomic features such as genes. This results in various name space conflicts, where different identifiers are linked to the same feature, corresponding features are not linked between name spaces, or the definition of the feature differs between name spaces. Studies have found a high level of discrepancy among the mapping resources, where querying a mapping resource with e.g. a microarray probe set identifier can result in a widely different list of related proteins or Gene Ontology terms (43,44). Challenges also include changes in the identifiers between different data versions. Various solutions such as the Life Science Identifier (LSID) have been created to solve these issues, but the field is yet to embrace these standards (45,46). One of the more succesful genomic identifier standardization endeavours includes the work done by The HUGO Gene Nomenclature Committee (HGNC). HGNC has curated and assigned unique gene symbols and names to over 33,000 human loci, enabling clear and unambiguous referencing of genes and

therefore also facilitating electronic data retrieval from databases and publications based on gene symbols and names (47). Another major attempt for genomic information standardization is the work conducted by the Locus Reference Genomic (LRG) project. LRG is collaboration between the two major sequence providers (EBI and NCBI) and various other genomic databases and research laboratories. The LRG collaboration aims to provide stable and unique references to human genomic sequences to be used as a reference standard for reporting disease-causing variants in human genes (48).

Unmatched data types. It is not always clear how biological entities such as DNA sequences, genes, mRNA, regulatory elements, proteins, functional and structural attributes are related to each other. These relations can be one-to-one relationships (e.g. protein and originating gene), as well as one-to-many (e.g. gene to several produced proteins). They can be incomplete (e.g. gene expression microarray probe representing part of transcribed mRNA or orthologous genes between species) or based on probability (e.g. predicted interaction). In addition, it is not always clear if and how the data could be compared, for example in a situation where you have different types of data, such as quantitative value of gene expression, and a discrete genomic locus representing a binding site. Many tools and methods have been developed for performing this type of integrative data analysis and fusion but many challenges remain (49–55).

Data visualization. Integration of information from different sources also increases the complexity of the data. This in turn complicates the interpretation and exploration of the data. Therefore methods and tools are required for visualizing the data, thus making it possible for humans to more easily digest and process all of the integrated information. Various visualization tools have been developed towards this end, and these include network-based tools such as Cytoscape (56), and genome centered tools such as the Integrative Genomics Viewer (IGV) (57) and the UCSC Genome Browser (58).

Some of these challenges will be discussed in greater detail later in this thesis during the presentation of the novel solutions presented in the original publications.

4 Aims

The main aims of the work presented in this thesis were to develop new methods and tools for integrating and interpreting genomic data.

The specific aims were:

- To study existing methods and software tools available for integrating genomic data (Publication I).

- To develop a novel method and software tool for integrating heterogeneous cross-species and cross-platform genomic data sets (Publication II).

- To develop a novel method and software tool for visualizing integrated genomic data sets using three-dimensional force-directed graph networks (Publication III).

- To develop a human genetic variation database portal that integrates data from various genomic databases thus facilitating easy inspection of large sets of genetic variations (Publication IV).

5 Materials and methods

This chapter describes the different methods applied in the original publications and also provides a list of the different data sources used throughout the studies.

5.1 CROSS-LINKING AND METAGENE INTEGRATION

Cross-linking is a naïve method of integrating heterogeneous data, and in its simplicity consists of linking different identifiers to each other. A popular example is linking a gene to the products of this gene (i.e. transcripts and proteins). Cross-linking is often used to convert identifiers from one identifier system to another, for example converting microarray gene expression probe set identifiers to gene symbols.

The main advantage of cross-linking integration is its simplicity, while its main disadvantage is lost accuracy. This is evident when converting from identifiers to more general identifiers, for example when cross-linking a transcript identifier to a gene symbol. If the gene produces, as an example, ten different transcripts, after the conversion to a gene symbol, it is no longer possible to distinguish between the different transcripts, and therefore cannot convert the gene symbol back to the same transcript identifier. Publication I contains a more detailed explanation about cross-link integration, and a comparison of publicly available on-line cross-linking tools.

CROPPER tool presented in the Publication II implements a variation of the cross-linking integration, termed metagene integration. The concept of metagene integration is based on creating a new conceptual gene called a “metagene”, which is then cross-linked to corresponding genes in different species, and through the genes, to different gene products and identifiers. In a sense, all the different orthologs of a gene and their gene products are collapsed into a single metagene (Figure 2). Being able to convert different genomic identifiers to universal metagene identifiers greatly simplifies the task of integrating heterogeneous data, though the lost accuracy is a major concern similarly to other cross-link integration methods.

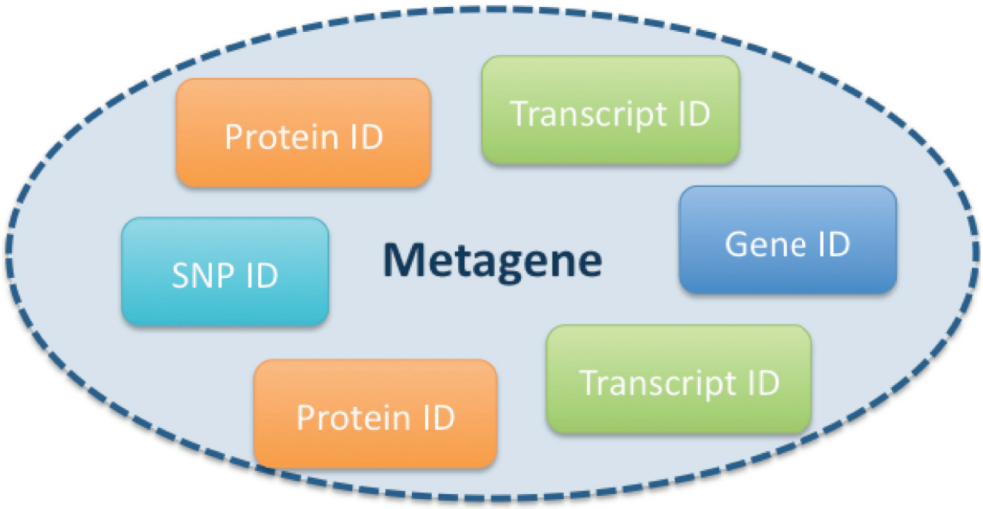

Figure 2. Conceptual illustration of metagenes. Essentially all related genomic identifiers are collapsed into a single metagene identifier, resulting in ease of integration and loss of accuracy.

5.2 DATA-ANALYSIS METHODS

This section briefly describes the various data-analysis and statistical methods applied in the original publications.

A standard score (z-score) method was used to standardize values from heterogeneous data sets prior to analysis in Publication II. The z-score indicates how many standard deviations an observation is above or below the mean, and is therefore a useful method when combining values from heterogeneous sources where the scale and distributions differ. Standardized z-score is calculated by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. (59)

Self-organizing map (SOM) is a type of artificial neural network that can be used to organize samples to a low-dimensional topological representation. A self-organizing map consists of nodes, which during a training phase are trained to represent distinctive elements of the training data. When populating the map, samples are distributed in the map so that they are close to nodes similar to themselves. This will result in an arrangement of the samples where similar samples are grouped together, facilitating cluster analysis. This type of analysis has been successfully

used with gene expression microarray data (60,61), and was used to cluster samples in Publication II.

Gene set enrichment analysis (GSEA) is a group of methods where a set or ranked list of genes is inspected for over- or under-representation of given themes when compared to a background set of genes, from where the gene set or list was originally drawn (e.g., are more genes linked to a certain pathway than you would expect by a random chance) (62,63). GSEA analysis is often performed when researchers want to inspect if their gene set/list obtained through an experiment is enriched with regards to different pathways, biological functions, chromosomal locations or diseases. In Publication II, differentially expressed genes were studied for enrichment based on Gene Ontology (GO) terms and KEGG pathways using DAVID (64) and GENERATOR (65) tools.

5.3 FORCE-DIRECTED GRAPHS

Force-directed graphs are an intuitive way of visualizing network graphs in an aesthetically pleasing manner. In addition to visual properties, force-directed graphs allow assignment of physical properties to the nodes and edges of the graph (66). In the software tool FORG3D presented in Publication III, editable visual attributes include the color, size and shape of the nodes, and color, width and direction of the edges, while physical attributes include mass and electric constant for nodes, and spring constant for edges. The edges are modeled as springs, and nodes as electrically charged particles. This allows modeling of the network as a physical system, where nodes and edges are modeled with basic laws of physics (Hooke's law for edges and Coulomb's law for nodes). The forces are applied to the nodes, pushing them farther away, while the edges constrain their movement.

Coulomb's law states that the electrostatic force between two charged particles (nodes) can be presented as $F_c = \frac{k_e q_1 q_2}{r^2}$, and the restoring force of a spring (edge) based on Hooke's law can be presented as $F_h = -k_s kr$, where r is the distance between the nodes, q_1 and q_2 are their charges, k is the spring constant of the connecting edge, while k_e and k_s are the global electric and spring constants. When the simulation is running, the particles try to achieve a distance where these forces are in equilibrium,

this distance can be presented as $r = \sqrt[3]{\frac{k_e q_1 q_2}{-k_s k}}$. The global damping constant representing friction is subtracted from the forces using the following formula $F=-kv$, where $-k$ is the global damping constant and v is the velocity. The simulation works by taking the forces based on Coulomb's and Hooke's law and assigning them to Newton's law of motion $F=ma$. Newton's Laws enables relating the position, velocity and acceleration of the simulated nodes as a differential equation for the unknown position of the node as a function of time. Numerical integration can therefore be used to solve the differential equation and advance the simulation by a given time step.

In FORG3D, this simulation is visualized in three-dimension in real-time allowing users to see how the different forces apply to the forming of the network, and also interact with the simulation to see how moving, editing or removing nodes or edges affects the formation and behavior of the network.

5.4 SOFTWARE DEVELOPMENT TOOLS

Different software development tools and programming languages were used when performing the studies and developing the related software.

Perl programming language (67) was used for facilitating text-file management, parsing and merging in Publications III and IV. Perl was also used for the implementation of the CROPPER software (Publication II). CROPPER's web user-interface is a Perl based Common Gateway Interface (CGI) software that interacts with a Perl based backend. The backend utilizes Bioperl (68) and Ensembl Perl API (69) packages for handling the biological information and retrieving data from a local installation of the Ensembl database (70).

MySQL relational database management system (71) was used in the CROPPER and Varietas software tools (Publications II and IV). CROPPER utilized a local copy of the Ensembl database (70), while Varietas utilized a custom integrated database called VarietasDB.

PHP programming language (72) was used for implementing the web user-interface of the Varietas tool (Publication IV). Varietas' user-interface retrieves data from the VarietasDB database based on the user's queries.

R Project for Statistical Computing (73) was used in Publication IV for retrieving and combining data from the Ensembl Biomart database using biomaRt (74) Bioconductor (75) package.

GCC C++ compiler (76) was used for developing the FORG3D software (Publication III). OpenGL graphics library (77) was used for implementing the graphical user interface (GUI) and 3D visualization.

5.5 DATA SOURCES

The developed software tools integrate and utilize data from a variety of public data sources, which are briefly described in Table 3. These data sources are utilized in various ways throughout the original publications, commonly as initial sources of data or as external resources for further information about selected parts of the data.

Table 3. List of used data sources.

Name	Description
ArrayExpress	Database of functional genomic experiments
Ensembl	Genome database
Gene Ontology (GO)	Database of gene and gene product attributes
Genetic Association Database (GAD)	Database of human genetic association studies
NCBI Entrez Gene	Gene database
NCBI Entrez SNP	Database of small genomic variations
NCBI GEO	Gene expression database
NCBI OMIM	Catalog of human genes and genetic disorders
NCBI Pubmed	Biomedical literature citation database
NHGRI GWAS Catalog	Catalog of Published Genome-Wide Association Studies
SNPedia	Wiki for human SNP information
WikiGenes	Wiki for gene information
WormBase	*C. elegans* genome database

6 Results

This chapter describes results from the original publications presented in this thesis, while Figure 3 illustrates how these studies relate to common steps in analysis of integrated data. Publications I, II and IV focus on combining data, Publication III focuses on visualization of combined data, while Publications III and IV also provide methods and tools for interpreting the results from the data-analysis. Original publications II and III also contain brief exemplary data analyses.

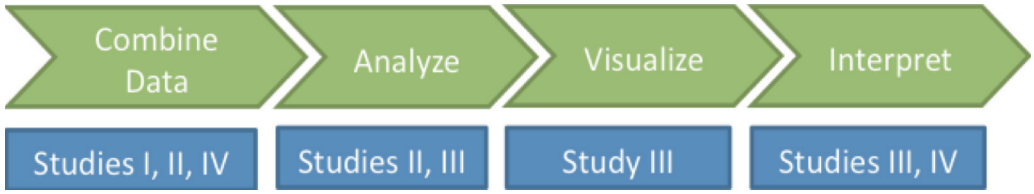

Figure 3. Common steps in analysis of integrated data and how the presented studies relate to these steps. In practice these steps are usually iterated several times during an analysis.

6.1 EXISTING DATA INTEGRATION SOFTWARE TOOLS

The work presented in Publication I lays out the groundwork for the other studies included in this thesis. The main reason for initiating this study was to get an overview of the available genomic data integration software tools, and to summarize this information for a wider audience, as well as use it as the basis for planning of our own software tools.

The publication describes the different methods, standards and tools for genomic data integration, especially focusing on cross-link integration where different types of genomic identifiers are cross-linked to each other. Besides introducing the available methods, standards and tools, the study reviews eight on-line data integration tools (64,78–84) and compares them based on criteria such as usability, cross-species and cross-platform functionality, result processing and availability of

different identifiers and data types. These on-line integration tools were selected based on PubMed literature searches, and only publicly available on-line resources utilizing cross-link integration were selected for review. The study also highlights shortcomings and weaknesses of the available tools, and suggests how these tools could be improved. One of these suggestions is an emphasis on the standardization of how genomic information is stored and referenced, and the suggested design for these standards would include model of the data, controlled vocabulary for describing the data, and XML-based markup language created from the model and vocabulary.

Based on the study, the following features were found to be desirable in a genomic data integration tool:

- Good coverage of data (e.g. supported species, technologies and types of genomic information)
- Intuitive and friendly user-interface
- Regular updates of data content
- Reliable availability of the service
- Batch processing of large amounts of data
- Possibility to input data in various formats
- Possibility to customize results output
- Possibility to preview the results in a web browser and download them as a separate file
- Application programming interface (API) and/or direct database access for using the tool programmatically

The main result of the study is the realization that the field lacked a tool that would be usable for integration of genomic data from wide variety of research technologies and species, and the list of desirable features for such a tool.

The publication has been referenced various times, especially as an example of bioinformatics approach used in pharmacological research (85–88).

Results from this study were used when designing and implementing new methods and tools represented in Publications II-IV.

6.2 CROSS-SPECIES AND -PLATFORM DATA INTEGRATION

Publication II describes a method and software tool for cross-linking integration using metagenes. The rationale for developing the metagene method and the associated CROPPER tool was to create an easily usable tool that would facilitate combining cross-species data from experiments conducted on various microarray gene expression platforms. The lessons learned while working with Publication I about desirable data integration tool features were taken into account when designing and implementing the software. During the course of development the scope of the tool was expanded to also allow integration of other types of genomic information, such as DNA sequences and proteins. One of the major design goals was to enable integration of publicly available gene expression microarray data sets to new data produced by our own experiments.

Metagenes are identifiers that group together all different identifiers linked to a single gene and its products, as well as orthologous genes and their products in different species. The software tool, CROPPER, provides an easy and quick way of performing cross-linking of heterogeneous identifiers, such as different gene and gene product identifiers from different technologies, databases and species to common metagene identifiers. It then facilitates combining of these metagene identifiers, further enabling the performance of integrative cross-platform and cross-species studies (Figure 4).

The original publication also includes a brief exemplary analysis, where Parkinson's disease data from different heterogeneous sources such as different technology platforms (protein array and gene expression microarrays) and species (human, mouse, *C. elegans* and macaque) are combined. The analysis grouped the data into 4,055 common metagenes, of which 247 were differentially expressed in the human Parkinson disease data and in the animal Parkinson disease model data sets. The analysis showcases how the metagene approach can be used to identify groups of co-regulated genes and proteins in Parkinson's disease and disease models, and provides a tool for hypothesis creation and further validation of analysis results by combining data from publicly available data sets.

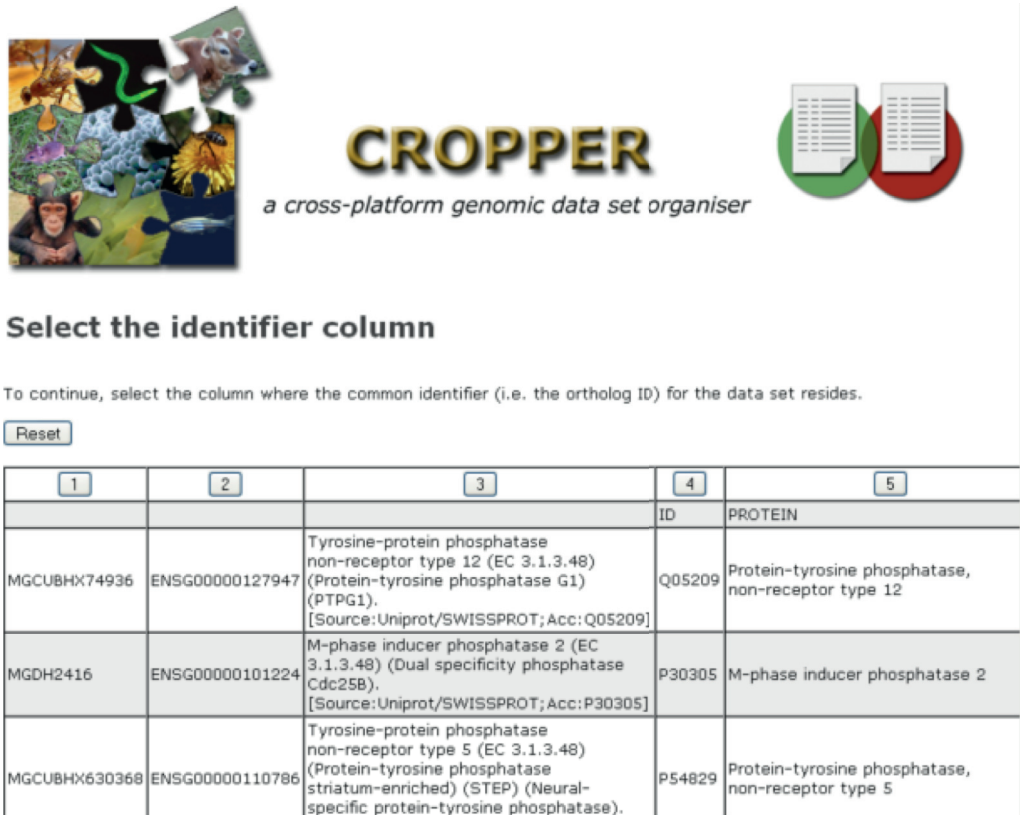

Figure 4. Screenshot of CROPPER. Metagene identifiers (first column) have been created for each row in the original dataset allowing using the column as a common key for combining the dataset with others.

The main result from the study is the development of the metagene integration algorithm, described in the Materials and Methods section, and the implementation of this algorithm as a part of the user-friendly CROPPER webtool. In our own use the tool has been successfully used for facilitating various studies combining cross-species data, especially microarray gene expression data from *C. elegans* and human experiments.

The publication has also been cited as an example of cross-species integration tool (89,90).

6.3 VISUALIZATION OF INTEGRATED GENOMIC DATA

Visualization is one of the major challenges when working with integrated genomic data, mainly because the large amount of data and its

complex structure makes interpretation of the data a demanding task. Different visualizations help researchers to understand and explore these structures and to get a better grasp of what different parts of the data actually mean the in the context of the original research questions. For integrated genomic data, visualization of network graphs might be one of the most important and useful areas of visualization. The reason for this is that many phenomena in genomics can be represented as graphs, for example different signaling pathways such as gene and protein interactions (56). In addition almost any kind of data set can be represented as a graph, because correlations, associations, and different distance metrics can be interpreted as graphs.

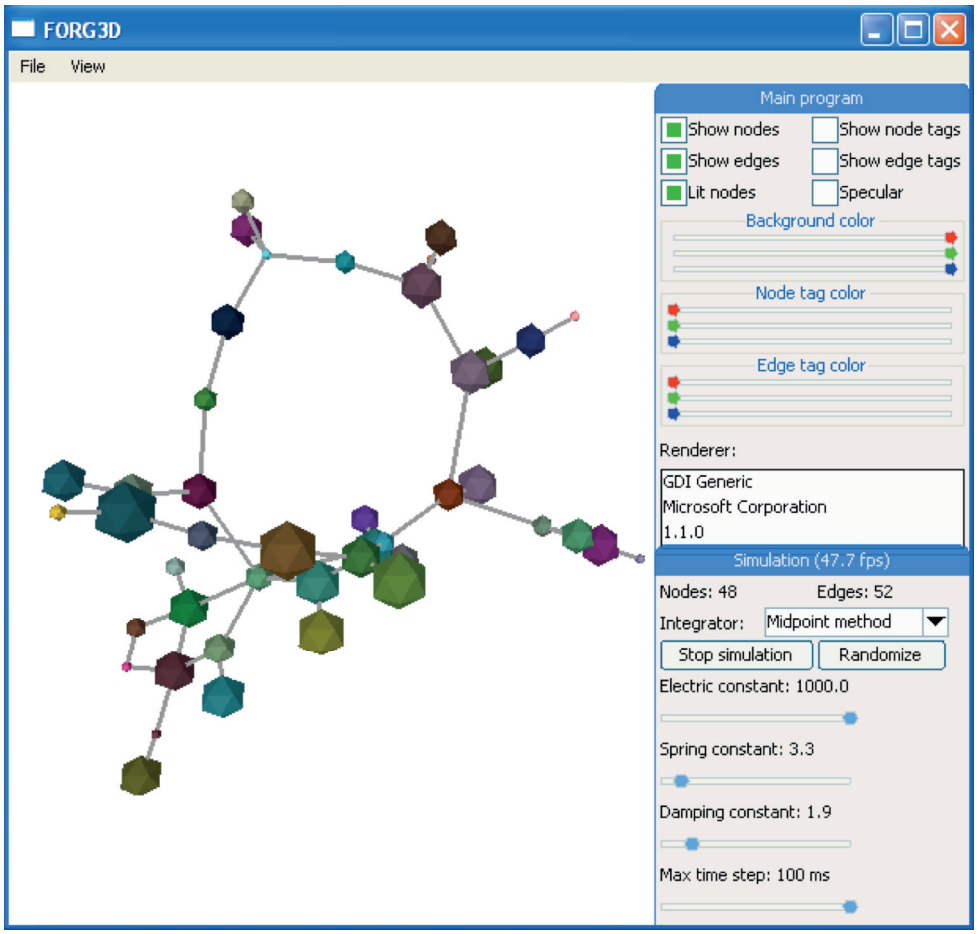

Figure 5. User-interface of FORG3D, including a network containing nodes and edges of different shapes, colors and physical properties.

To help researchers visualize integrated genomic data, a method and software tool was developed for visualization of three-dimensional force-directed graphs. The software, called FORG3D, allows researchers to create, edit and explore these network graphs, while visualizing them in real-time in three dimensions. Figure 5 shows a screenshot from FORG3D, displaying the user-interface, as well as how different visual properties can be assigned to the nodes and edges of a network.

In addition to developing the method and software, a small case study was performed, where *C. elegans* genetic interaction data was integrated with microarray gene expression data (Parkinson's disease model worms compared to wild type worms) and Gene Ontology (GO) data, and then visualized and explored with FORG3D. The resulting network consisted of 449 nodes representing genes and 1,223 edges representing interactions between the genes. The color of the nodes was set to represent gene expression fold-change, while the size of the node represented number of associated GO annotations. Width of the edges represented interaction strength and the color represented the type of evaluation method for the interaction. Besides demonstrating the concept of visualizing integrated data with FORG3D, one of the main results from the analysis was identifying *daf-2* gene as one of the main hubs of the differentially expressed genes, thus indicating the potential importance of *daf-2* in regulating gene expression in this Parkinson's disease model. *daf-2* is an insulin/growth factor receptor, which has been in several studies linked to aging (91,92).

The main result of the project is the implementation of physics based force-directed 3D networks as a standalone desktop software that can easily handle real-time interaction with large (up to hundreds of thousands of nodes and edges) networks. The original motivation for developing FORG3D was to have a readily usable tool that would allow interactive and swift real-time visualization of tens-of-thousands of genes from compendium gene expression microarray studies that combine data from public and private data sets, and FORG3D has been successfully used for this purpose in various projects. In addition, FORG3D has been used to visualize other types of network data, including social media networks derived from sources such as Internet Relay Chat (IRC), Facebook and Twitter.

The publication has been cited various times as an example of network visualization tool (93–98).

6.4 DATABASE PORTAL OF HUMAN GENETIC VARIATION

Genetic variations such as SNPs, copy number variants (CNVs) and insertions/deletions are of a major interest for researchers studying the genetic components of different traits and diseases (6,99,100). SNP microarrays and sequencing technologies enable researchers to inspect millions of different genetic variants in selected populations, while data-analysis methods can be used to find associations between these variations and phenotypes of interest, often resulting in a large number of potential associations (101,102). Browsing through these candidate associations and linking them to genes or other functional elements, and retrieving information about these variations, related genes and functional elements is important, but also time consuming and challenging.

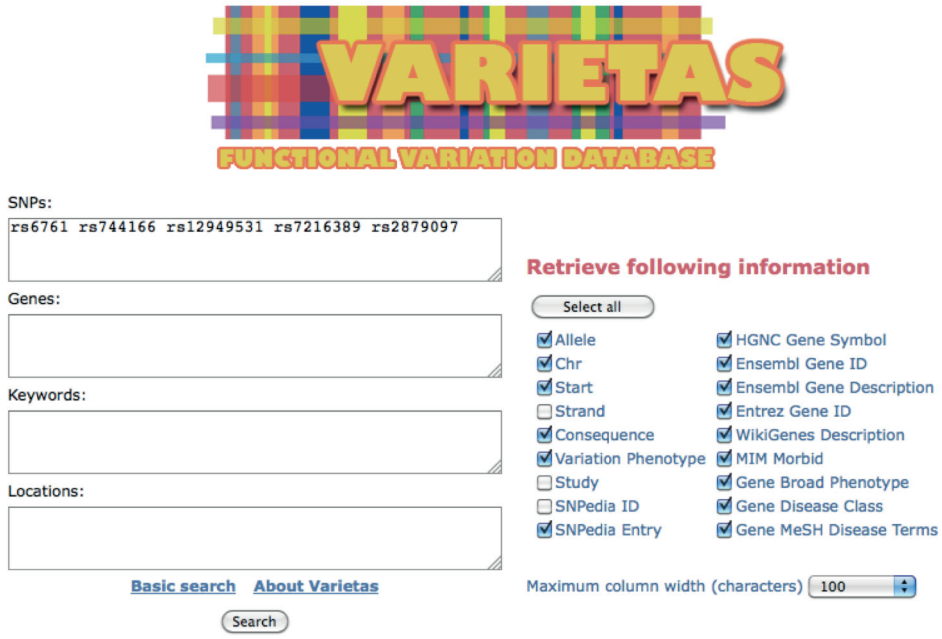

Figure 6. Screenshot of Varietas. Varietas allows users to query information based on variation or gene identifiers, keywords and genomic locations. Various types of information linked to the query terms can be retrieved.

To facilitate this process, a web-based database portal was developed, thus allowing researches to easily retrieve information about large sets of

variations, related genes, genomic elements, phenotypes and diseases. The database portal, called Varietas, integrates and links out to a variety of genome and variation databases and resources (Figure 6).

An example data set consisting of five SNPs (rs6761, rs744166, rs12949531, rs7216389 rs2879097) is provided to demonstrate the database and Varietas provides annotations for these variations from Ensembl, SNPedia, NHGRI GWAS Catalog and Genetic Association Database (GAD) and links them to corresponding genes and gene information from Ensembl, NCBI Entrez Gene, NCBI OMIM and NCBI PubMed.

The main result of this project is the Varietas database that integrates data from these various external sources, as well as the easy web-user interface that can be used to query the database.

Varietas has been successfully used in a large number of internal projects where candidate SNPs from candidate gene and GWAS studies are investigated, ranked and annotated (103–107). Externally Varietas website is accessed monthly by hundreds of unique users performing various types of database queries. Varietas has also been featured as a Tip of the Week in the OpenHelix blog where the OpenHelix team has provided a video tutorial of how Varietas can be used for genetic research (available at: http://blog.openhelix.eu/?p=5287). The publication has been cited as an example of a variation database (108,109) as well as a tool used for assessing biological roles of variations identified in candidate gene studies (96).

6.5 RESULTS SUMMARY

The main results from the original publications are the evaluation of existing cross-linking integration methods and tools (Publication I), and the developed novel algorithms and the implemented software tools CROPPER (Publication II), FORG3D (Publication III) and Varietas (Publication IV).

The original publications and their supplementary materials contain more detailed information about the tools including: benchmarking compared to other available tools, availability of the tools, links to the software, descriptions of software architecture and implementation, as well as detailed methods and results from the exemplary data-analyses performed in Publications II and III.

7 Discussion

This chapter discusses the theoretical and practical implications of the work presented in this thesis, describes the limitations of the research and developed methods and tools, and also takes a look in to the emerging challenges and possible solutions.

7.1 THEORETICAL AND PRACTICAL IMPLICATIONS

Much work remains to be done in the field of genomic information integration as the field expands and new types of data appear at an increasing pace. The work presented in this thesis describes a small part of this effort, while also introducing new methods and tools to help researchers to integrate genomic information.

Even a great methodological idea can be useless if it is difficult to apply in practice, and therefore a special emphasis has been placed on usability and accessibility of the developed tools. The tools have been well received and have already contributed to the field in the form of being used in research and in teaching. A good example is the Varietas database (Publication IV), which has attracted thousands of users since its release and has been featured and linked to from various sites focused on genetic research.

Besides developing new methods and implementing tools based on them, the feasibility of data integration and visualization was demonstrated through case studies in publications II and III. These case studies are far from actual full-sized compendium studies, but showcase how the tools and ideas can be applied in practice.

7.2 LIMITATIONS AND CONSIDERATIONS

The work presented in this thesis is limited in many ways. It only focuses on a small part of genomic data integration, and does not attempt to provide solutions or an in-depth look into aspects such as statistical or software engineering methods used in data integration. The tools

described in original publications II-IV did not directly discuss metabolomic, epigenetic or next-generation sequencing data, even though for example FORG3D tool (Publication III) can be used to visualize these types of data as well, and Varietas database contains data about human genetic variations obtained, among others technologies, by next-generation sequencing.

The availability of developed and published web-resources is an important issue, and unfortunately the tool CROPPER, (Publication II) is an example of a published web-tool that is no longer supported or accessible because of lack of funding for required server infrastructure and maintenance. CROPPER was available on-line from 2004 to 2009. It is extremely important to plan in advance how the software tools are deployed, maintained, updated and supported, including plans for infrastructure, personnel and funding. Fortunately these issues are increasingly discussed by scientific journals and funding agencies (110).

One of the practical limitations concerning the usability of FORG3D is the lack of support for common formats used for biological network data, such as Graph Modeling Language (GML), BioPAX biological pathway (111) or Proteomics Standards Initiative Molecular Interaction (PSI-MI) (112) formats. This limitation clearly demonstrates the benefit of supporting a standard format, while also underlining the difficulty of selecting supported formats when a large number of different standards exist.

7.3 FUTURE PERSPECTIVES

The future of genomic information looks to be driven by the appearance of new types of data, as well as the ever-increasing amount of produced data. This progress is likely to continue and will emphasize the need for new standards, methods and tools for genomic data management and analysis. The focus has already shifted from a gene-centric view to a gene set-centric, and is likely to shift to genome and genome set-centric views, practically meaning comparing, combining and analyzing together complete genomes. The new areas facing the challenges of genomic data integration also includes the area of high-content analysis (HCA) where new imaging technologies are producing massive amounts of data from biological experiments (113).

The increasing amount of data also means that there will be more and more ways of combining, analyzing and interpreting the data. There already are endless possibilities for performing integrative analyses, and therefore more focus needs to be placed on deciding the best experimental designs and most interesting research questions. In short, this is a challenge that cannot be solved simply by adding more computational resources, but needs to be addressed by smart researchers who understand the underlying biology and analytical methods.

Effort should be placed on creating new widely accepted standards for structuring and sharing data. This is also an intricate question, as it is not obvious who should develop these standards (e.g. research groups, research consortiums, private or public companies, large data centers such as NCBI or EBI), and what are the best ways of getting users to adopt them. Examples of successful and widely used standards are Gene Ontology project (standardized representation of gene and gene product attributes) (114), Minimum Information About a Microarray Experiment - MIAME (115) and sequence formats such as FASTA and FASTQ (116). Examples of newly introduced data formats include Variant Call Format (VCF) that is taking its place as the leading format for managing genomic variation data called from sequencing data, and SeqXML and OrthoXML that have been developed for storing sequence and orthologue data in a structured format (117).

An interesting development is also the progress done in the field of information management, not necessarily focusing on life sciences. These developments include creation of the Semantic Web, aimed at structuring data available on the World Wide Web (WWW). As a part of this effort, a family of specifications called The Resource Description Framework (RDF) have been developed, and later applied also to the field of biosciences, for example in the Bio2RDF project (118). This is a good example of how progress done in the field of biomedical data management and integration will also depend on progress happening in the field of information technology and data management.

8 Summary

This thesis describes current progress in the field of genomic data integration, and provides new insights, methods and tools for scientists working in biomedical research. The original publications in this thesis are as follows:

Publication I, includes a review and comparison of existing genomic data integration tools, while also discussing the different methods used for cross-linking genomic data identifiers.

Publication II, introduces a new metagene based approach and web-tool for combining data for heterogeneous cross-platform and cross-species compendium studies, the publication also demonstrates combining Parkinson's disease datasets from four separate organisms, that were produced with different research technologies.

Publication III, focuses on the visualization of integrated genomic data through force-directed 3D graphs, and describes a software tool that can be used for this type of visualization, the publication also includes a case study where different types of genomic data from individual *C. elegans* studies are integrated and visualized, showcasing the potential of force-directed graph visualization for explorative data analysis.

Publication IV, describes a new web-based database that integrates data from various genomic databases, and allows large-scale retrieval of information about human genetic variations such as SNPs, indels and copy number variations.

The thesis discusses these publications and the field as a whole, while also casting a look in to the future perspectives of genomic data integration.

References

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004 Oct 21;431(7011):931–45.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.
3. Johannsen W. The Genotype Conception of Heredity. The American Naturalist. 1911;45(531):129–59.
4. Houle D, Govindaraju DR, Omholt S. Phenomics: the next challenge. Nature Reviews Genetics. 2010 Dec;11(12):855–66.
5. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007 Oct 18;449(7164):851–61.
6. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep;467(7311):52–8.
7. International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec 18;426(6968):789–96.
8. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics. 2010 Oct;42(11):937–48.
9. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010 Aug;466(7307):707–13.
10. Durbin RM, Altshuler DL, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061–73.
11. Su W-L, Kleinhanz RR, Schadt EE. Characterizing the role of miRNAs within gene regulatory networks using integrative genomics techniques. Molecular Systems Biology. 2011 May 24;7.

12. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nature reviews. Genetics. 2010 Dec;
13. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011 May 15;advance on.
14. Wass MN, Fuentes G, Pons C, Pazos F, Valencia A. Towards the prediction of protein interaction partners using physical docking. Molecular Systems Biology. 2011 Feb 15;7.
15. Cohen AA, Geva-Zatorsky N, Eden E, Frenkel-Morgenstern M, Issaeva I, Sigal A, et al. Dynamic proteomics of individual cancer cells in response to a drug. Science (New York, N.Y.). 2008 Dec 5;322(5907):1511–6.
16. Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, et al. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nature biotechnology. 2010 Jan;28(1):83–9.
17. Nygren H, Seppänen-Laakso T, Castillo S, Hyötyläinen T, Orešič M. Liquid chromatography-mass spectrometry (LC-MS)-based lipidomics for studies of body fluids and tissues. Methods in molecular biology (Clifton, N.J.). 2011 Jan;708:247–57.
18. Soininen P, Kangas AJ, Würtz P, Tukiainen T, Tynkkynen T, Laatikainen R, et al. High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. The Analyst. 2009 Sep;134(9):1781–5.
19. Hu C, van der Heijden R, Wang M, van der Greef J, Hankemeier T, Xu G. Analytical strategies in lipidomics and applications in disease biomarker discovery. Journal of chromatography. B, Analytical technologies in the biomedical and life sciences. 2009 Sep 15;877(26):2836–46.
20. Wymann MP, Schneiter R. Lipid signalling in disease. Nature reviews. Molecular cell biology. 2008 Feb;9(2):162–76.
21. Wheelock CE, Goto S, Yetukuri L, D'Alexandri FL, Klukas C, Schreiber F, et al. Bioinformatics strategies for the analysis of lipids. Methods in molecular biology (Clifton, N.J.). 2009 Jan;580:339–68.
22. Lankinen M, Schwab U, Gopalacharyulu PV, Seppänen-Laakso T, Yetukuri L, Sysi-Aho M, et al. Dietary carbohydrate modification alters serum metabolic profiles in individuals with the metabolic syndrome.

Nutrition, metabolism, and cardiovascular diseases : NMCD. 2010 May;20(4):249–57.

23. Lankinen M, Schwab U, Seppänen-Laakso T, Mattila I, Juntunen K, Mykkänen H, et al. Metabolomic analysis of plasma metabolites that may mediate effects of rye bread on satiety and weight maintenance in postmenopausal women. The Journal of nutrition. 2011 Jan;141(1):31–6.
24. Carone BR, Fauquier L, Habib N, Shea JM, Hart CE, Li R, et al. Paternally Induced Transgenerational Environmental Reprogramming of Metabolic Gene Expression in Mammals. Cell. 2010 Dec 23;143(7):1084–96.
25. Ng S-F, Lin RCY, Laybutt DR, Barres R, Owens JA, Morris MJ. Chronic high-fat diet in fathers programs β-cell dysfunction in female rat offspring. Nature. 2010 Oct;467(7318):963–6.
26. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. American journal of human genetics. 2011 Apr 8;88(4):450–7.
27. Beisel C, Paro R. Silencing chromatin: comparing modes and mechanisms. Nature Reviews Genetics. 2011 Jan 11;12(2):123–35.
28. Djupedal I, Ekwall K. Epigenetics: heterochromatin meets RNAi. Cell research. 2009 Mar;19(3):282–95.
29. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, N.Y.). 2009 Oct 9;326(5950):289–93.
30. Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. Determinants of nucleosome organization in primary human cells. Nature. 2011 May 22;advance on.
31. Zhang Z, Wippo CJ, Wal M, Ward E, Korber P, Pugh BF. A Packing Mechanism for Nucleosome Organization Reconstituted Across a Eukaryotic Genome. Science. 2011 May 19;332(6032):977–80.
32. Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, et al. A map of open chromatin in human pancreatic islets. Nature genetics. 2010 Mar;42(3):255–9.
33. Small KS, Hedman ÅK, Grundberg E, Nica AC, Thorleifsson G, Kong A, et al. Identification of an imprinted master trans regulator at the

KLF14 locus related to multiple metabolic phenotypes. Nature Genetics. 2011 May 15;43(6):561–4.
34. Montgomery SB, Dermitzakis ET. From expression QTLs to personalized transcriptomics. Nature Reviews Genetics. 2011 Mar;advance on.
35. Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, et al. Defining the transcriptome and proteome in three functionally different human cell lines. Molecular systems biology. 2010 Dec 21;6:450.
36. Ghazalpour A, Bennett B, Petyuk VA, Orozco L, Hagopian R, Mungrue IN, et al. Comparative Analysis of Proteome and Transcriptome Variation in Mouse. PLoS Genetics. 2011 Jun 9;7(6):e1001393.
37. Inouye M, Kettunen J, Soininen P, Silander K, Ripatti S, Kumpula LS, et al. Metabonomic, transcriptomic, and genomic variation of a population cohort. Molecular Systems Biology. 2010 Dec 21;6:441.
38. Kitano H. Systems biology: a brief overview. Science (New York, N.Y.). 2002 Mar 1;295(5560):1662–4.
39. Kilpinen S, Autio R, Ojala K, Iljin K, Bucher E, Sara H, et al. Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. Genome biology. 2008 Jan;9(9):R139.
40. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic acids research. 2010 Nov 21;39(suppl_1):D1005–10.
41. Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, et al. ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic acids research. 2010 Nov 10;39(Database issue):D1002–4.
42. Kahn SD. On the future of genomic data. Science (New York, N.Y.). 2011 Feb 11;331(6018):728–9.
43. Drăghici S, Sellamuthu S, Khatri P. Babel's tower revisited: a universal resource for cross-referencing across annotation databases. Bioinformatics (Oxford, England). 2006 Dec 1;22(23):2934–9.
44. Day R, McDade K, Chandran U, Lisovich A, Conrads T, Hood B, et al. Identifier mapping performance for integrating transcriptomics and proteomics experimental results. BMC Bioinformatics. 2011;12(1):213.

45. Clark T, Martin S, Liefeld T. Globally distributed object identification for biological knowledgebases. Briefings in bioinformatics. 2004 Mar;5(1):59–70.
46. Martin S, Hohman MM, Liefeld T. The impact of Life Science Identifier on informatics data. Drug discovery today. 2005 Nov 15;10(22):1566–72.
47. Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. genenames.org: the HGNC resources in 2011. Nucleic acids research. 2011 Jan;39(Database issue):D514–9.
48. Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, et al. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome medicine. 2010 Jan;2(4):24.
49. Lê Cao K-A, González I, Déjean S. integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics (Oxford, England). 2009 Nov 1;25(21):2855–6.
50. Huopaniemi I, Suvitaival T, Nikkila J, Oresic M, Kaski S. Multivariate multi-way analysis of multi-source data. Bioinformatics. 2010 Jun 6;26(12):i391–8.
51. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical applications in genetics and molecular biology. 2009 Jan;8(1):Article28.
52. Lê Cao K-A, Martin PGP, Robert-Granié C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. BMC bioinformatics. 2009 Jan;10(1):34.
53. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nature Reviews Genetics. 2010 Jun;11(7):476–86.
54. Quackenbush J. Data reporting standards: making the things we use better. Genome medicine. 2009 Jan;1(11):111.
55. Chervitz SA, Deutsch EW, Field D, Parkinson H, Quackenbush J, Rocca-Serra P, et al. Data standards for Omics data: the basis of data sharing and reuse. Methods in molecular biology (Clifton, N.J.). 2011 Jan;719:31–69.
56. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, et al. Integration of biological networks and gene expression data using Cytoscape. Nature protocols. 2007 Jan;2(10):2366–82.

57. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology. 2011 Jan;29(1):24–6.
58. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome research. 2002 Jun;12(6):996–1006.
59. Cheadle C, Cho-Chung YS, Becker KG, Vawter MP. Application of z-score transformation to Affymetrix data. Applied bioinformatics. 2003 Jan;2(4):209–17.
60. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences of the United States of America. 1999 Mar;96(6):2907–12.
61. Törönen P, Kolehmainen M, Wong G, Castrén E. Analysis of gene expression data using self-organizing maps. FEBS letters. 1999 May;451(2):142–6.
62. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005 Oct;102(43):15545–50.
63. Kurki MI, Paananen J, Storvik M, Yla-Herttuala S, Jaaskelainen JE, von Und Zu Fraunberg M, et al. TAFFEL: Independent Enrichment Analysis of gene sets. BMC bioinformatics. 2011 May 19;12(1):171.
64. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome biology. 2003 Jan;4(5):P3.
65. Pehkonen P, Wong G, Törönen P. Theme discovery from gene lists for identification and viewing of multiple functional groups. BMC bioinformatics. 2005 Jan;6:162.
66. Fruchterman TMJ, Reingold EM. Graph drawing by force-directed placement. Software: Practice and Experience. 1991 Nov;21(11):1129–64.
67. Perl. The Perl Programming Language [Internet]. 2011 [cited 2010 Nov 9];Available from: http://www.perl.org/

68. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, et al. The Bioperl toolkit: Perl modules for the life sciences. Genome research. 2002 Oct;12(10):1611–8.
69. Stabenau A, McVicker G, Melsopp C, Proctor G, Clamp M, Birney E. The Ensembl core software libraries. Genome research. 2004 May;14(5):929–33.
70. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, et al. Ensembl 2011. Nucleic acids research. 2010 Nov;
71. MySQL. MySQL database [Internet]. 2011 [cited 2010 Nov 11];Available from: http://www.mysql.com/
72. PHP. PHP: Hypertext Preprocessor [Internet]. 2011 [cited 2010 Nov 11];Available from: http://www.php.net
73. R Development Core Team. R: A Language and Environment for Statistical Computing. 2010;
74. Durinck S, Huber W. biomaRt: Interface to BioMart databases [Internet]. 2010;Available from: http://bioconductor.org/packages/release/bioc/html/biomaRt.html
75. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome biology. 2004 Jan;5(10):R80.
76. GCC. GCC, the GNU Compiler Collection [Internet]. 2011;Available from: http://gcc.gnu.org/
77. OpenGL. The OpenGL Graphics System [Internet]. 2011;Available from: http://www.opengl.org/
78. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic acids research. 2003 Jan;31(1):219–23.
79. Tsai J, Sultana R, Lee Y, Pertea G, Karamycheva S, Antonescu V, et al. RESOURCERER: a database for annotating and linking microarray resources within and across species. Genome biology. 2001 Jan;2(11):SOFTWARE0002.
80. Wang P, Ding F, Chiang H, Thompson RC, Watson SJ, Meng F. ProbeMatchDB--a web database for finding equivalent probes across microarray platforms and species. Bioinformatics (Oxford, England). 2002 Mar;18(3):488–9.

81. Bussey KJ, Kane D, Sunshine M, Narasimhan S, Nishizuka S, Reinhold WC, et al. MatchMiner: a tool for batch navigation among gene and gene product identifiers. Genome biology. 2003 Jan;4(4):R27.
82. Cheung K-H, Hager J, Pan D, Srivastava R, Mane S, Li Y, et al. KARMA: a web server application for comparing and annotating heterogeneous microarray platforms. Nucleic acids research. 2004 Jul;32(Web Server issue):W441–4.
83. Liefeld T, Reich M, Gould J, Zhang P, Tamayo P, Mesirov JP. GeneCruiser: a web service for the annotation of microarray data. Bioinformatics (Oxford, England). 2005 Sep;21(18):3681–2.
84. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, et al. EnsMart: a generic system for fast and flexible access to biological data. Genome research. 2004 Jan;14(1):160–9.
85. Hammerling U, Tallsjö A, Grafström R, Ilbäck N-G. Comparative hazard characterization in food toxicology. Critical reviews in food science and nutrition. 2009 Aug;49(7):626–69.
86. Whittaker PA. Can pharmacology possibly have a role for bioinformatics? Expert Opinion on Drug Discovery. 2007 Feb 16;2(2):271–84.
87. Neubauer M, Ross-Macdonald P. Annual Reports in Medicinal Chemistry Volume 42. Elsevier; 2007.
88. Chang X. Bayesian Mixtures and Gene Expression Profiling with Missing Data. 2008.
89. Fierro AC, Vandenbussche F, Engelen K, Van de Peer Y, Marchal K. Meta Analysis of Gene Expression Data within and Across Species. Current genomics. 2008 Dec;9(8):525–34.
90. Kuhn A, Luthi-Carter R, Delorenzi M. Cross-species and cross-platform gene expression studies with the Bioconductor-compliant R package "annotationTools". BMC bioinformatics. 2008 Jan;9:26.
91. Gami MS, Wolkow CA. Studies of Caenorhabditis elegans DAF-2/insulin signaling reveal targets for pharmacological manipulation of lifespan. Aging cell. 2006 Feb;5(1):31–7.
92. Kenyon C, Chang J, Gensch E, Rudner A, Tabtiang R. A C. elegans mutant that lives twice as long as wild type. Nature. 1993 Dec;366(6454):461–4.

93. Lucas M, Laplaze L, Bennett MJ. Plant systems biology: network matters. Plant, cell & environment. 2011 Apr;34(4):535–53.
94. Blank LM, Kuepfer L. Metabolic flux distributions: genetic information, computational predictions, and experimental validation. Applied microbiology and biotechnology. 2010 May 1;86(5):1243–55.
95. Campbell SJ, Gaulton A, Marshall J, Bichko D, Martin S, Brouwer C, et al. Visualizing the drug target landscape. Drug discovery today. 2010 Jan;15(1-2):3–15.
96. Kelsey RM, Alpert BS, Dahmer MK, Krushkal J, Quasney MW. Alpha-adrenergic receptor gene polymorphisms and cardiovascular reactivity to stress in Black adolescents and young adults. Psychophysiology. 2012 Mar;49(3):401–12.
97. Brown KR, Otasek D, Ali M, McGuffin MJ, Xie W, Devani B, et al. NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. Bioinformatics (Oxford, England). 2009 Dec 15;25(24):3327–9.
98. Fucile G, Di Biase D, Nahal H, La G, Khodabandeh S, Chen Y, et al. ePlant and the 3D data display initiative: integrative systems biology on the world wide web. PloS one. 2011 Jan;6(1):e15237.
99. Durbin RM, Altshuler DL, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct;467(7319):1061–73.
100. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. Nature reviews. Genetics. 2010 May;11(5):356–66.
101. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics. 2011 May 8;advance on.
102. Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nature genetics. 2010 Nov;42(11):949–60.
103. Koutnikova H, Laakso M, Lu L, Combe R, Paananen J, Kuulasmaa T, et al. Identification of the UBP1 locus as a critical blood pressure determinant using a combination of mouse and human genetics. PLoS genetics. 2009 Aug;5(8):e1000591.

104. Stancáková A, Kuulasmaa T, Paananen J, Jackson AU, Bonnycastle LL, Collins FS, et al. Association of 18 confirmed susceptibility loci for type 2 diabetes with indices of insulin release, proinsulin conversion, and insulin sensitivity in 5,327 nondiabetic Finnish men. Diabetes. 2009 Sep;58(9):2129–36.
105. Vangipurapu J, Stančáková A, Pihlajamäki J, Kuulasmaa TM, Kuulasmaa T, Paananen J, et al. Association of indices of liver and adipocyte insulin resistance with 19 confirmed susceptibility loci for type 2 diabetes in 6,733 non-diabetic Finnish men. Diabetologia. 2010 Dec;
106. Stancakova A, Paananen J, Soininen P, Kangas AJ, Bonnycastle LL, Morken MA, et al. Effects of 34 Risk Loci for Type 2 Diabetes or Hyperglycemia on Lipoprotein Subclasses and Their Composition in 6,580 Nondiabetic Finnish Men. Diabetes. 2011 Mar;:db10–1655-.
107. Kaarniranta K, Paananen J, Nevalainen T, Sorri I, Seitsonen S, Immonen I, et al. Adiponectin receptor 1 gene (ADIPOR1) variant is associated with advanced age-related macular degeneration in Finnish population. Neuroscience Letters. 2012 Feb;
108. Glusman G, Caballero J, Mauldin D, Hood L, Roach JC. KAVIAR: an accessible system for testing SNV novelty. Bioinformatics (Oxford, England). 2011 Sep 28;27(22):3216–7.
109. Li MJ, Wang P, Liu X, Lim EL, Wang Z, Yeager M, et al. GWASdb: a database for human genetic variants identified by genome-wide association studies. Nucleic acids research. 2012 Jan;40(Database issue):D1047–54.
110. Chandras C, Weaver T, Zouberakis M, Smedley D, Schughart K, Rosenthal N, et al. Models for financial sustainability of biological databases and resources. Database : the journal of biological databases and curation. 2009 Jan 23;2009(0):bap017.
111. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, et al. The BioPAX community standard for pathway data sharing. Nature Biotechnology. 2010 Sep 9;28(9):935–42.
112. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, et al. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nature biotechnology. 2004 Feb;22(2):177–83.

113. Collins MA. Generating “omic knowledge”: the role of informatics in high content screening. Combinatorial chemistry & high throughput screening. 2009 Nov;12(9):917–25.
114. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000 May;25(1):25–9.
115. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nature genetics. 2001 Dec 1;29(4):365–71.
116. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research. 2010 Apr;38(6):1767–71.
117. Schmitt T, Messina DN, Schreiber F, Sonnhammer ELL. SeqXML and OrthoXML: standards for sequence and orthology information. Briefings in Bioinformatics. 2011 Jun 11;:bbr025-.
118. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics. 2008 Oct;41(5):706–16.

ORIGINAL PUBLICATIONS (I-IV)

I

Integration of genomic data for pharmacology and toxicology using Internet resources

Jussi Paananen, Garry Wong

SAR QSAR Environ Res. 2006 Feb;17(1):25-36

SAR and QSAR in Environmental Research
Vol. 17, No. 1, February 2006, 25–36

Taylor & Francis
Taylor & Francis Group

Integration of genomic data for pharmacology and toxicology using Internet resources¶

J. PAANANEN† and G. WONG*‡§

†Department of Computer Science, University of Kuopio, Finland
‡Department of Biochemistry, University of Kuopio, Finland
§A. I. Virtanen Institute for Molecular Sciences, University of Kuopio, P.O. Box 1627, 70211 Kuopio, Finland

(*Received 31 October 2005; in final form 16 December 2005*)

Genome based technologies such as sequencing and gene expression profiling using microarrays are creating massive amounts of data. Results from these studies have provided unique insights into targets, biochemical pathways, and biological systems affected by drug or xenobiotic chemical treatments. Moreover, these genomic technologies offer the potential to identify biomarkers for pharmacological development or toxicological prediction. Nonetheless, microarray studies involving a single compound produce useful although limited data. To gain further power from these individual studies, the ability to combine datasets through integration schemes has become imperative. In the current study, we describe and analyze currently available Internet resources designed to address this problem. Many functionalities, such as ability to cross reference orthologous genes across species or to combine same technology platform data, are present in these resources. Nonetheless, these resources are limited in the number of technology platforms they can support. While the ability to integrate all currently existing gene expression datasets remains enigmatic, the current tools provide a partial solution that may still yield unique insights into the affects of exogenous molecules at the level of gene expression.

Keywords: Genomic data; Integration; Bioinformatics; Internet

1. Introduction

Genome based research methods provide researchers with enormous amounts of experimental data such as from sequencing, gene expression profiling, protein arrays and many other high-throughput approaches. In many cases this experimental data is collected and saved in specialized databases. Compendium studies are then used to combine data from these different experiments [1–3]. These studies usually focus on a certain biological question and combine data from different experiments.

*Corresponding author. Email: garry.wong@uku.fi
¶Presented at CMTPI 2005: Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (Shanghai, China, October 29–November 1, 2005).

SAR and QSAR in Environmental Research
ISSN 1062-936X print: ISSN 1029-046X online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/10659360600562053

An example of pharmacology based compendium study would be to determine the effect of a drug treatment on gene expression using a single agent with different time points, doses, and routes of administration. The study could be extended across species by adding experimental data from the same drug treatment to different animal models such as mice or rats. Compendium studies have the promise of revealing new drug-gene associations, while at the same time being cost-efficient since they use the available information from previous experiments.

One of the most crucial parts of compendium studies is combining and integrating the available data. This is the most difficult step and it is typical that the data cannot be directly combined due to biological or technological issues. The biological reasons include the lack of available corresponding data, for example gene expression datasets may be collected from the same drug treatment in mice and rats, but the dose, time point, and tissue sampled may be different. The technical challenges are caused by factors related to presenting and storing experimental data. It is common that various research equipment and software systems process and structure the data in different ways, therefore complicating integration of the data.

Extensive and well structured information is required for reliable integration of experimental data. At the moment several standards, ontologies, and object models are being developed towards this end. The purpose of standards is to provide a common framework for data management, while ontologies provide controlled explicit vocabulary that can be used to describe data. Object models, on the other hand, are used to describe relationships between different parts of data and can be used to design databases and information systems. The best known genomic data standards and ontologies include Gene Ontology [4] that provides a vocabulary to describe gene and gene product attributes, and works of Microarray Gene Expression Data Society (MGED) [5]. These include Minimum Information About a Microarray Experiment (MIAME) standard [6] and Microarray And Gene Expression (MAGE) object model and markup language [7]. MGED also has workgroups working on the fields of data transformation and normalization, *in situ* hybridization and immunohistochemistry, nutrigenomics, environmental genomics and toxicogenomics. These standards will only grow to be more essential for integration of genomic data as they mature.

Different solutions and methods have been developed for integrating genomic data and these methods work in two different levels. Higher level integration methods integrate whole information systems and databases, while lower level of integration focuses only in integration of data. Higher level integration where different resources are combined uses common software engineering approaches such as middleware architectures where a new middleware information system is created to work as a mediator between existing information systems. Another common approach is data warehousing where data is collected from different databases and stored in one massive data warehouse; for example Ensembl database [8] uses data warehousing methods to combine data from several specialized databases, such as Rat Genome Database [9], FlyBase [10], Wormbase [11] and *Saccharomyces* Genome Database [12]. Benefits of higher level integration methods are usually good performance combined with increased functionality. On the other hand, this kind of integration requires a great deal of work and resources, close collaboration with administrators of integrated resources and extended maintenance. Even with higher level integration there usually is also a need for lower level integration to actually combine the data within the integrated resources. Because higher level integration is arduous and resource demanding, many solutions

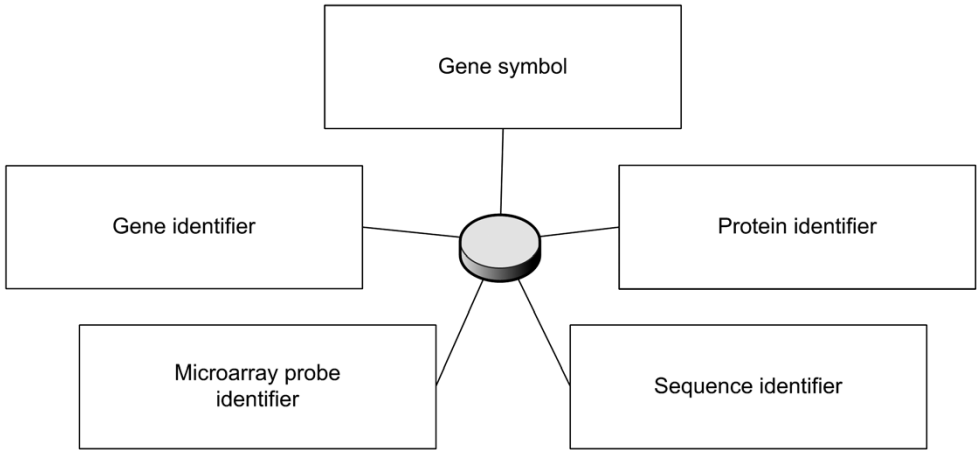

Figure 1. A diagram for cross-linking. In cross-linking, data identifiers are mapped to identifiers from different sources. These sources can be technology or database identifiers or species specific identifiers that can be mapped between species based on orthologue information. The nodes represent different identifiers and the central hub idealises how these identifiers can be linked with one another.

have only focused on integration of data. The simplest and therefore most used method for combining data is so called ''cross-linking/link-integration''. Cross-linking is based on mapping data identifiers to corresponding identifiers from different sources (see figure 1).

The reliability of cross-linking is highly dependent on how the mapping has originally been done. Many sophisticated techniques have been developed towards this end [13–15] and these techniques use sequence analysis, statistical and other methods for linking data, but also more direct methods are used. The limitation of these methods is loss of accuracy, for example when linking a short sequence to a complete gene or putative protein. Because new sequencing and analysis information is produced constantly and the information stored to genomic databases is changing regularly, updating cross-linking information is important. This can be achieved by redoing the mapping based on updated information, but when referring to large genomic databases and sophisticated cross-linking methods, the actual linking may require weeks of computing time even with modern computing facilities. Therefore it is typical that the actual cross-linking of identifiers is done occasionally and information about the linked identifiers is then saved on a database where it can be accessed without the need of performing any resource consuming computing to find the corresponding identifiers. Because of this, it is not unusual that cross-linking information can be out of date and erroneous. Despite these severe limitations, cross-linking is still a popular way of integrating data from heterogeneous sources.

Many databases and technology providers have developed their own system for indexing and identifying genomic data. A lot of genomic resources have their own unique data identifiers that have to be mapped to identifiers in other resources. This mapping is often performed by databases and technology providers that want to increase usability of their services. Some of the best known genomic data resources are European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database [16], Ensembl, UniProt/SwissProt [17], Gene Ontology and National Center of

Biotechnology Information (NCBI) resources such as UniGene, GenBank, LocusLink, OMIM, Entrez, RefSeq and PubMed [18–20]. Many integration methods rely on existing data mapping between these databases provided by the databases themselves. There also exists studies on the efficiency of linkage between different databases [21].

Different resources have been developed for automated integration of genomic datasets and many of these resources are based on cross-linking of data identifiers. In the present study we describe, analyse, and compare the currently available Internet resources for combining genome based datasets, especially microarray or gene expression profiling sets. The purpose is to demonstrate the different approaches, feasibility, and implementations used towards integrating genome based datasets.

2. Methods

Different genomic data integration resource publications were searched using PubMed literature database. These publications were inspected and evaluated. Only on-line resources publicly available on the Internet using cross-link integration methods were selected. Evaluation of resources was based on the resource, publication, manuals and other supplementary material found. In testing the actual resource, emphasis was placed on usability and functionality of the resource. Different tests were carried out by using available test datasets from the websites of resources, or when not available, generating a random test dataset based on the specifications found from the publications or websites.

Based on tests of the resource and other available material, the following were studied: possibility to manage full datasets including experimental measurements; possibility to automatically merge distinct datasets; flexibility with input dataset structure; control over structure of result dataset; origin of the cross-link mapping data; species covered; overall usability and any additional features or information that separated the resource from its competitors.

In addition to the reviewed resources, two other published resources were inspected and found to be discontinued [22, 23].

3. Results

3.1 *DAVID*

DAVID [24] is a database for annotation, visualisation and integrated discovery, developed by National Institute of Allergy and Infectious Diseases, National Institute of Health, Bethesda, MD, USA. Identifier mapping and annotation information is derived from LocusLink and other NCBI's databases. Affymetrix [25], LocusLink, UniGene, UniProt, RefSeq and GenBank identifiers can be linked to each other using DAVID. Cross-species mapping information is obtained from HomoloGene database [18]. DAVID also contains tools for analysis of data and allows users to upload their own datasets that besides data identifiers contain experimental measurements. DAVID includes a wide variety of functionality for processing and analysing data, but structure of imported datasets is strictly predetermined and users also have limited control over structure of result files. Result files can be viewed and downloaded using

a web browser. DAVID development team has also announced that application programming interfaces (APIs) for developing software that can be integrated with DAVID will be published shortly. This would allow users to create their own custom programs for managing experimental data using tools from DAVID.

3.2 *EnsMart*

EnsMart [26] is part of the Ensembl database based on BioMart platform [27] and is used for batch processing and data mining of the Ensembl database. Ensembl is developed by EMBL-European Bioinformatics Institute (EBI), Hinxton, Cambridge, UK and the Wellcome Trust Sanger Institute (WTSI), Hinxton, Cambridge, UK. EnsMart can be used to retrieve data from Ensembl database and therefore can be used to cross-link data identifiers between different database identifiers and species available in Ensembl. These identifiers cover most of major database identifiers and species such as human, mouse rat, chimp, dog, cattle, chicken, zebrafish, fruitfly, mosquito, honey bee, yeast, rhesus macaque, elephant, opossum, *Xenopus tropicalis*, *Tetraodon nigroviridis*, *Ciona intestianalis*, *Ciona savignyi* and *Caenorhabditis elegans*. EnsMart offers a very attractive choice for cross-linking data. Because Ensmart is directly linked to Ensembl database, it covers a wide variety of information useful for cross-linking data across different platforms and species. EnsMart is also very user friendly and offers a good control over structure of result files. The result files can be viewed and downloaded using a web browser or delivered through e-mail. The main limitation of EnsMart is that is not actually meant for processing and combining datasets, but rather processing single lists of data identifiers. The actual integration of separate datasets has to be facilitated using other methods. Ensembl also offers a direct database connection to Ensembl and EnsMart databases, as well as a software toolkit [28] that can be used to access these databases. Therefore users are able to create their own custom software for the actual integration process.

3.3 *GeneCruiser*

GeneCruiser [29] is a web service for annotating and mapping of microarray data, developed by the Broad Institute of MIT and Harvard, Cambridge, MA, USA. Identifier mapping and annotation information is obtained periodically from SwissProt, TIGR human and mouse gene indices [30], UniGene, LocusLink and RefSeq databases. GeneCruiser also allows linking of data to corresponding information in the University of California Santa Cruz (UCSC) Genome Browser [31], PubMed, GenBank, Gene Cards [32] and the Gene Expression Omnibus [33]. Species covered by GeneCruiser are human, mouse, rat, yeast, zebrafish, fruitfly and *Arabidopsis thaliana*. GeneCruiser uses some novel approaches, such as Life Science Identifier (LSID) [34] developed by the Object Management Group (OMG). The LSID is a standard developed to facilitate management of different identifiers used in life sciences, and is especially useful for version control of identifiers. Besides a web interface that can be used to process a list of identifiers and to view or download result files, GeneCruiser is available as a Web service. GeneCruiser offers a public SOAP Web service interface defined in Web Services Description Language (WSDL). This Web service interface can be used for easy integration of GeneCruiser to other information systems, allowing users to create their own software that uses GeneCruiser functionality. Example of this kind of integration is incorporating of GeneCruiser to analysis platform called GenePattern.

The main limitation of GeneCruiser is that it is solely focused on Affymetrix microarrays, meaning that you are only able to cross-link identifiers to or from Affymetrix probe identifiers.

3.4 *KARMA*

KARMA [35] is a web server application for comparing and annotating heterogeneous microarray platforms, developed by research groups in Yale University, New Haven, USA and Celera, Rockville, MD, USA. KARMA can be used to cross-link and combine data from several predefined microarray resources, but also enables users to upload their own custom datasets. Identifier mapping and annotation information is obtained periodically from UniGene and cross-species information from HomoloGene databases. Data can be cross-linked to LocusLink, SwissProt, Ensembl and Gene Ontology data identifiers. Species available in KARMA are human, mouse, rat and *Arabidopsis thaliana*. Because KARMA allows users to upload and use their own custom datasets, KARMA can be useful in a wide variety of situations where datasets are obtained from different sources, but even with custom datasets users have very limited control over the structure of the result files. The result files are delivered through e-mail and can not be viewed or downloaded using a web browser. Because many e-mail servers have size limitations for file attachments, this may result in problems when dealing with large datasets. The main limitation of KARMA is that it requires datasets to include UniGene identifiers, as these identifiers are used for cross-linking. Therefore if the dataset to be processed does not contain UniGene identifiers, different software is first required to map available identifiers to UniGene identifiers, making KARMA in many cases obsolete. Also the limited amount of species covered by KARMA narrows its usability.

3.5 *MatchMiner*

MatchMiner [36] is a tool for batch navigation among gene and gene product identifiers, developed by Genomics and Bioinformatics Group, National Cancer Institute, Bethesda, MD, USA and SRA International Inc., Fairfax, VA, USA. Mapping information is based on USCS, LocusLink, UniGene, OMIM and Affymetrix databases. Species covered by MatchMiner are human and mouse, but no cross-species mappings are available. A special feature of MatchMiner is batch merging functionality that allows users to input two distinct datasets for combining. Users have limited control over structure of result files. Result files can be viewed and downloaded using a web browser.

3.6 *ProbeMatchDB*

ProbeMatchDB [37] is a web database for finding equivalent probes across microarray platforms and species. It is developed by University of Michigan Medical School, Ann Arbor, MI, USA. Mapping information is based on UniGene and HomoloGene databases. ProbeMatchDB allows users to input a list of Affymetrix, cDNA array clone, SwissProt, SAGE, UniGene, or Ensembl identifiers and cross-link them to each other. Species covered by ProbeMatchDB are human, mouse and rat. ProbeMatchDB allows cross-species mapping between these species. ProbeMatchDB can only process lists of data identifiers and because of this the actual integration of separate

datasets has to be facilitated with using other methods. Users have limited control over structure of result files. Result files can be viewed and downloaded using a web browser.

3.7 *RESOURCERER*

RESOURCERER [38] is developed by The Institute of Genomic Research (TIGR), Rockville, MD, USA. RESOURCERER is a microarray-resource annotation and cross-reference database built using the analysis of expressed sequence tags (ESTs) and gene sequences provided by the TIGR Gene Indicis (TGI) and TIGR Orthologous Gene Alignment (TOGA) [39] databases. RESOURCERER database is updated every four months and it contains mapping information for several commonly available microarray resources. These resources include microarray datasets from major microarray manufacturers such as Affymetrix and Agilent. Covered species are human, rat, mouse, zebrafish, cattle, *Caenorhabditis elegans* and *Xenopus tropicalis*. These microarray resources are cross-linked to identifiers including UniGene, LocusLink, RefSeq, PubMed and GeneOntology database identifiers. Also cross-species information is available for some identifiers. Users have very limited control over the structure of the result files that can be accessed through a web browser and also downloadable from the web server. Even though RESOURCERER includes information about dozens of different microarray resources, it is still hindered by this limited list. Therefore using custom datasets is not possible and for example for most of the species only one or two different resources are available, greatly limiting usability of the program. RESOURCERER is only useful when you are working with microarray resources covered by it, and in most of cases there is very limited functionality when working with other species than human, rat or mouse.

3.8 *SOURCE*

SOURCE [40] is a unified genomic resource of functional annotations, ontologies and gene expression data developed by research groups at Stanford University, Stanford, CA, USA. SOURCE is based on information retrieved from databases such as UniGene, LocusLink and SwissProt. SOURCE accepts datasets consisting of list of GenBank, UniGene or LocusLink identifiers, which in addition to mentioned identifiers can be cross-linked to Gene Ontology and UniProt identifiers. Species covered by SOURCE are human, mouse and rat. There is no functionality available for cross-species linking of orthologous genes. SOURCE can be used to provide extensive information on single genes, but batch function used for processing datasets has much more limited functionality. The user also has very limited control over the structure of the result file which can be viewed and downloaded using a web browser.

4. Discussion

Many different resources are available for cross-link integration of genomic data (see table 1).

In most of these resources mapping of data identifiers within species are based on UniGene and cross-species mappings on HomoloGene databases. Therefore, the

Table 1. Summary of the reviewed resources.

Name	*Species*	*Cross-species mapping*	*Data allowed in processing*	*Result delivery*	*Other*	*Web address*
DAVID	Species covered by HomoloGene	Yes	Yes	View, download		http://david.abcc.ncifcrf.gov/
EnsMart	21 different species	Yes	No	View, download, e-mail	Attached to Ensembl database Software API available	http://www.ensembl.org/
GeneCruiser	human, mouse, rat, yeast, zebrafish, fruitfly, *Arabidopsis thaliana*	Yes	No	View, download	Web service for software integration	http://www.genecruiser.org/
KARMA	human, mouse, rat, *Arabidopsis thaliana*	Yes	Yes	E-mail		http://biryani.med.yale.edu/karma/
MatchMiner	human, mouse	No	No	View, download	Merging of distinct datasets	http://discover.nci.nih.gov/matchminer/
ProbeMatchDB	human, mouse, rat	Yes	No	View, download		http://brainarray.mhri.med.umich.edu/
RESOURCERER	human, rat, mouse, zebrafish, cattle, *Caenorhabditis elegans* and *Xenopus tropicalis*	Yes	No	View, download	Does not allow custom datasets	http://www.tigr.org/tigr-scripts/magic/r1.pl
SOURCE	human, mouse, rat	No	No	View, download		http://source.stanford.edu/

main difference between these resources is the type of functionality built upon mappings. Many of the resources are purely just identifier converters that take a list of identifiers which are then converted to other identifiers and possibly annotated. In these cases, if a user wants to combine two distinct datasets that contain different identifiers, the user is first required to process datasets so that they have common identifiers and then use another program for the actual integration. An exception to this rule is MatchMiner that can also do the actual combining of datasets. Overall, most of the resources gave the user basically none or very little control over the structure of the input and result files, also emphasising the need of using additional programs to re-structure result files for further analysis.

It should be noted that many of the resources are only usable if specific requirements are met, meaning that users are limited to work with certain species and technologies covered by the resource. The focus on most of the reviewed resources is also on gene expression data, and there definitely is a need for more generic tools that allow management and integration of data produced with different technologies, including proteomics technologies such as protein microarrays and datasets queried from protein databases. The one resource that is closest to reaching this goal is EnsMart. Because EnsMart is built upon Ensembl which incorporates data from several databases, such as NCBI's databases, UniProt/SwissProt and several specialised organism and technology databases, it can be very useful in many different cases. On the other hand, Ensembl is not directly meant for processing of actual datasets and therefore has very limited functionality beside converting and annotation identifiers. This limitation can be overcome by developing programs that perform this function directly or using provided software API connects to Ensembl database. Besides Ensembl, GeneCruiser and DAVID provide, or are planning to provide, software integration means for software developers to use functionality of these resources in their own programs. From these the most sophisticated software integration method is the Web service provided by GeneCruiser. GeneCruiser also shows a good example by using the LSID standard, but there still is need to actually implement many of the available standards when developing bioinformatical software resources.

In summary, while many tools currently exist for integration of genomic, especially gene expression based data for pharmacology and toxicology research, the lack of standards has proven to be a serious hindrance. As datasets from studies designed to understand the effects of chemical compounds on biological systems continues to increase at a rapid rate, future efforts should be more highly focussed on the ability to easily integrate these results. In addition to existing integration projects where genomic databases are integrated using data warehousing and middleware approaches there is a clear need for standards for genomic data management. These standards should consist of object model of the data, controlled vocabulary for describing the data, and markup language based on the model and vocabulary. Such standards have already successfully been developed for managing different types of biological [41, 42], medical [43–45], pharmacological [46], and chemical [47, 48] data, and the attractiveness of extensible markup language (XML) as a standard for structuring data within the context of bioinformatics integration has been well described [49]. Standards like these are on the forefront of data management, and developing and implementing them holds the promise of facilitating automated genomic data management and integration in the future.

Acknowledgements

The authors thank Jani Kekäläinen, Petri Pehkonen, Tom Kolmakow and Markus Storvik for discussions and helpful comments. The authors also wish to thank all the researchers who have taken part in making resources reviewed in this article available.

References

[1] S.K. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J.M. Stuart, A. Eizinger, B.N. Wylie, G.S. Davidson. *Science*, **293**, 2087 (2001).
[2] S. Bergmann, J. Ihmels, N. Barkai. *PLoS Biol*., **2**, 9 (2004).
[3] S.A. McCarroll, C.T. Murphy, S. Zou, S.D. Pletcher, C.S. Chin, Y.N. Jan, C. Kenyon, C.I. Bargmann, H. Li. *Nat. Genet*., **36**, 197 (2004).
[4] M.A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G.M. Rubin, J.A. Blake, C. Bult, M. Dolan, H. Drabkin, J.T. Eppig, D.P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J.M. Cherry, K.R. Christie, M.C. Costanzo, S.S. Dwight, S. Engel, D.G. Fisk, J.E. Hirschman, E.L. Hong, R.S. Nash, A. Sethuraman, C.L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S.Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E.M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White. *Nucleic Acids Res*., **32**, 258 (2004). Available online at: www.geneontology.org (accessed 1 December 2005).
[5] Microarray Gene Expression Data Society (MGED) website. Available online at: www.mged.org (accessed 1 December 2005).
[6] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, M. Vingron. *Nat. Genet*., **29**, 365 (2001). Available online at: www.mged.org (accessed 1 December 2005).
[7] P.T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W.L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B.J. Aronow, A. Robinson, D. Bassett, C.J. Stoeckert, Jr., A. Brazma. *Genome Biol*., **3**, (2002). Available online at: www.mged.org (accessed 1 December 2005).
[8] T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X.M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, E. Birney. *Nucleic Acids Res*., **33**, 447 (2005). Available online at: www.ensembl.org (accessed 1 December 2005).
[9] N. de la Cruz, S. Bromberg, D. Pasko, M. Shimoyama, S. Twigger, J. Chen, C.F. Chen, C. Fan, C. Foote, G.R. Gopinath, G. Harris, A. Hughes, Y. Ji, W. Jin, D. Li, J. Mathis, N. Nenasheva, J. Nie, R. Nigam, V. Petri, D. Reilly, W. Wang, W. Wu, A. Zuniga-Meyer, L. Zhao, A. Kwitek, P. Tonellato, H. Jacob. *Nucleic Acids Res*., **33**, 485 (2005). Available online at: http://rgd.mcw.edu/ (accessed 1 December 2005).
[10] R.A. Drysdale, M.A. Crosby. *Nucleic Acids Res*., **33**, 390 (2005). Available online at: www.flybase.org (accessed 1 December 2005).
[11] N. Chen, T.W. Harris, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, P. Canaran, J. Chan, C.K. Chen, W.J. Chen, F. Cunningham, P. Davis, E. Kenny, R. Kishore, D. Lawson, R. Lee, H.M. Muller, C. Nakamura, S. Pai, P. Ozersky, A. Petcherski, A. Rogers, A. Sabo, E.M. Schwarz, K. Van Auken, Q. Wang, R. Durbin, J. Spieth, P.W. Sternberg, L.D. Stein. *Nucleic Acids Res*., **33**, 383 (2005). Available online at: www.wormbase.org (accessed 1 December 2005).
[12] K.R. Christie, S. Weng, R. Balakrishnan, M.C. Costanzo, K. Dolinski, S.S. Dwight, S.R. Engel, B. Feierbach, D.G. Fisk, J.E. Hirschman, E.L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C.L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, D. Botstein, J.M. Cherry. *Nucleic Acids Res*., **32**, 311 (2004). Available online at: www.yeastgenome.org (accessed 1 December 2005).
[13] A.C. Culhane, G. Perriere, D.G. Higgins. *BMC Bioinformatics*, **4**, 59 (2003).
[14] P. Hu, C.M. Greenwood, J. Beyene. *BMC Bioinformatics*, **6**, 128 (2005).

[15] P. Warnat, R. Eils, B. Brors. *BMC Bioinformatics*, **6**, 265 (2005).
[16] C. Kanz, P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. van den Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F.G. Diez, N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, S. Sobhany, P. Stoehr, M.A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu, R. Apweiler. *Nucleic Acids Res*., **33**, 29 (2005). Available online at: www.ebi.ac.uk/embl/ (accessed 1 December 2005).
[17] A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.S. Yeh. *Nucleic Acids Res*., **33**, 154 (2005). Available online at: www.ebi.ac.uk/swissprot/ (accessed 1 December 2005).
[18] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, J.U. Pontius, K.D. Pruitt, G.D. Schuler, L.M. Schriml, E. Sequeira, S.T. Sherry, K. Sirotkin, G. Starchenko, T.O. Suzek, R. Tatusov, T.A. Tatusova, L. Wagner, E. Yaschenko. *Nucleic Acids Res*., **33**, 39 (2005). Available online at: www.ncbi.nlm.nih.gov (accessed 1 December 2005).
[19] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova. *Nucleic Acids Res*., **33**, 54 (2005). Available online at: www.ncbi.nlm.nih.gov (accessed 1 December 2005). Available online at: www.ncbi.nlm.nih.gov (accessed 1 December 2005).
[20] K.D. Pruitt, T. Tatusova, D.R. Maglott. *Nucleic Acids Res*., **33**, 501 (2005). Available online at: www.ncbi.nlm.nih.gov (accessed 1 December 2005).
[21] A. Veldhoven, D. de Lange, M. Smid, V. de Jager, J.A. Kors, G. Jenster. *BMC Bioinformatics*, **6**, 192 (2005).
[22] A.V. Kulkarni, N.S. Williams, Y. Lian, J.D. Wren, D. Mittelman, A. Pertsemlidis, H.R. Garner. *Bioinformatics*, **18**, 1410 (2002).
[23] B.A. Svensson, A.J. Kreeft, G.J. van Ommen, J.T. den Dunnen, J.M. Boer. *Genome Biol*., **4**, 35 (2003).
[24] G. Dennis, Jr., B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, R.A. Lempicki. *Genome Biol*., **4**, 3 (2003).
[25] G. Liu, A.E. Loraine, R. Shigeta, M. Cline, J. Cheng, V. Valmeekam, S. Sun, D. Kulp, M.A. Siani-Rose. *Nucleic Acids Res*., **31**, 82 (2003). Available online at: http://www.affymetrix.com/analysis/netaffx/ (accessed 1 December 2005).
[26] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, E. Birney. *Genome Res*., **14**, 160 (2004). Available online at: www.ensembl.org (accessed 1 December 2005).
[27] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, W. Huber. *Bioinformatics*, **21**, 3439 (2005).
[28] A. Stabenau, G. McVicker, C. Melsopp, G. Proctor, M. Clamp, E. Birney. *Genome Res*., **14**, 929 (2004).
[29] T. Liefeld, M. Reich, J. Gould, P. Zhang, P. Tamayo, J.P. Mesirov. *Bioinformatics*, **21**, 3681 (2005).
[30] Y. Lee, J. Tsai, S. Sunkara, S. Karamycheva, G. Pertea, R. Sultana, V. Antonescu, A. Chan, F. Cheung, J. Quackenbush. *Nucleic Acids Res*., **33**, 71 (2005). Available online at: www.tigr.org/tdb/tgi/ (accessed 1 December 2005).
[31] F. Hsu, T.H. Pringle, R.M. Kuhn, D. Karolchik, M. Diekhans, D. Haussler, W.J. Kent. *Nucleic Acids Res*., **33**, 454 (2005). Available online at: http://genome.ucsc.edu/ (accessed 1 December 2005).
[32] M. Safran, I. Solomon, O. Shmueli, M. Lapidot, S. Shen-Orr, A. Adato, U. Ben-Dor, N. Esterman, N. Rosen, I. Peter, T. Olender, V. Chalifa-Caspi, D. Lancet. *Bioinformatics*, **18**, 1542 (2002). Available online at: www.genecards.org (accessed 1 December 2005).
[33] R. Edgar, M. Domrachev, A.E. Lash. *Nucleic Acids Res*., **30**, 207 (2002). Available online at: www.ncbi.nlm.nih.gov/geo/ (accessed 1 December 2005).
[34] T. Clark, S. Martin, T. Liefeld. *Brief Bioinform*., **5**, 59 (2004).
[35] K.H. Cheung, J. Hager, D. Pan, R. Srivastava, S. Mane, Y. Li, P. Miller, K.R. Williams. *Nucleic Acids Res*., **32**, 441 (2004).
[36] K.J. Bussey, D. Kane, M. Sunshine, S. Narasimhan, S. Nishizuka, W.C. Reinhold, B. Zeeberg, W. Ajay, J.N. Weinstein. *Genome Biol*., **4**, 27 (2003).
[37] P. Wang, F. Ding, H. Chiang, R.C. Thompson, S.J. Watson, F. Meng. *Bioinformatics*, **18**, 488 (2002).
[38] J. Tsai, R. Sultana, Y. Lee, G. Pertea, S. Karamycheva, V. Antonescu, J. Cho, B. Parvizi, F. Cheung, J. Quackenbush. *Genome Biol*., **2** (2001).
[39] Y. Lee, R. Sultana, G. Pertea, J. Cho, S. Karamycheva, J. Tsai, B. Parvizi, F. Cheung, V. Antonescu, J. White, I. Holt, F. Liang, J. Quackenbush. *Genome Res*., **12**, 493 (2002). Available online at: www.tigr.org/tdb/tgi/ego/ (accessed 1 December 2005).
[40] M. Diehn, G. Sherlock, G. Binkley, H. Jin, J.C. Matese, T. Hernandez-Boussard, C.A. Rees, J.M. Cherry, D. Botstein, P.O. Brown, A.A. Alizadeh. *Nucleic Acids Res*., **31**, 219 (2003).
[41] A. Finney, M. Hucka. *Biochem. Soc. Trans*., **31**, 1472 (2003).
[42] J.J. Berman, *Hum. Pathol*., **36**, 139 (2005).
[43] J. Guo, A. Takada, K. Tanaka, J. Sato, M. Suzuki, T. Suzuki, Y. Nakashima, K. Araki, H Yoshihara. *J. Med. Syst*., **28**, 523 (2004).
[44] H. Wang, F. Azuaje, B. Jung, N. Black. *BMC Med. Inform. Decis. Mak*., **3**, 4 (2003).

[45] H. Sugimori, K. Yoshida, S. Hara, K. Furumi, I. Tofukuji, T. Kubodera, T. Yoda, M. Kawai. *Meth. Inf. Med.*, **41**, 220 (2002).
[46] D. Hanisch, R. Zimmer, T. Lengauer. *In Silico Biol.*, **2**, 313 (2002).
[47] Y.M. Liao, H. Ghanadan. *Anal. Chem.*, **74**, 389 (2002).
[48] N. Kikuchi, A. Kameyama, S. Nakaya, H. Ito, T. Sato, T. Shikanai, Y. Takahashi, H. Narimatsu. *Bioinformatics*, **21**, 1717 (2005).
[49] F. Achard, G. Vaysseix, E. Barillot. *Bioinformatics*, **17**, 115 (2001).

II

CROPPER: a metagene creator resource for cross-platform and cross-species compendium studies

Jussi Paananen, Markus Storvik, Garry Wong

BMC Bioinformatics. 2006 Sep 22;7:418.

BMC Bioinformatics

BioMed Central

Software

Open Access

CROPPER: a metagene creator resource for cross-platform and cross-species compendium studies

Jussi Paananen[1,2], Markus Storvik[3] and Garry Wong*[1,3]

Address: [1]Department of Neurobiology, A. I. Virtanen Institute for Molecular Sciences, P.O. Box 1627, 70211 Kuopio, Finland, [2]Department of Computer Science, University of Kuopio, P.O. Box 1627, 70211 Kuopio, Finland and [3]Department of Biochemistry, University of Kuopio, P.O. Box 1627, 70211 Kuopio, Finland

Email: Jussi Paananen - Jussi.Paananen@uku.fi; Markus Storvik - Markus.Storvik@uku.fi; Garry Wong* - Garry.Wong@uku.fi

* Corresponding author

Published: 22 September 2006

BMC Bioinformatics 2006, **7**:418 doi:10.1186/1471-2105-7-418

Received: 02 June 2006

Accepted: 22 September 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/418

Abstract

Background: Current genomic research methods provide researchers with enormous amounts of data. Combining data from different high-throughput research technologies commonly available in biological databases can lead to novel findings and increase research efficiency. However, combining data from different heterogeneous sources is often a very arduous task. These sources can be different microarray technology platforms, genomic databases, or experiments performed on various species. Our aim was to develop a software program that could facilitate the combining of data from heterogeneous sources, and thus allow researchers to perform genomic cross-platform/cross-species studies and to use existing experimental data for compendium studies.

Results: We have developed a web-based software resource, called CROPPER that uses the latest genomic information concerning different data identifiers and orthologous genes from the Ensembl database. CROPPER can be used to combine genomic data from different heterogeneous sources, allowing researchers to perform cross-platform/cross-species compendium studies without the need for complex computational tools or the requirement of setting up one's own in-house database. We also present an example of a simple cross-platform/cross-species compendium study based on publicly available Parkinson's disease data derived from different sources.

Conclusion: CROPPER is a user-friendly and freely available web-based software resource that can be successfully used for cross-species/cross-platform compendium studies.

Background

Novel genomic research methods have enabled researchers to perform high-throughput experiments that results in massive amounts of experimental data. Combining data from different experiments allows researchers to validate their results and to gain a better understanding of the biological questions being studied. In cross-species studies, data derived from experiments performed on different organisms are combined to find universal themes. In cross-platform studies, common biological questions are studied using different research platforms and technologies. The ability to combine experimental data is particularly useful when extended to combine data available on public data repositories such as sequence, expression, and literature databases.

The desire to perform large-scale studies that combine experimental results from various sources has led to com-

pendium studies that combine cross-species/cross-platform data in order to obtain a larger perspective on biological questions. Unfortunately, for several reasons, the combining of experimental results is anything but a trivial task. These reasons can be divided into biological and technical challenges. The biological challenges include variances between species, different experimental design/conditions, and lack of knowledge of the underlying biological processes. The technical challenges are caused by differences in how experimental data is stored, presented, and managed. Because of the lack of standards, different research equipment, software, and databases identify and structure data in different and unique ways that make it a challenge to combine data obtained from these heterogeneous sources.

To address these technical challenges and to enable researchers to automate the integration of the genomic data derived from heterogeneous sources, without the need for using complex programming and scripting tools, we have developed a user-friendly web-based software resource called CROPPER. CROPPER can be used to combine datasets from different genomic research platforms such as microarrays, biological databases, and experiments performed on different species. When performing the combining process, associated data can be brought along. This facilitates the import of the resulting dataset by the user into the desired statistical/analytical program for further analysis. CROPPER uses the latest genomic information with respect to identifiers and orthologous genes retrieved from the Ensembl [1] database.

Implementation

CROPPER is developed using Perl version 5.9.1, Bioperl version 1.4 [2] and Ensembl database API written in Perl [3]. CROPPER runs on a web-server which also acts as an application server. Users can use the web-interface to input their datasets and related parameters to the application server. The Application server processes the data, and if required, queries the database server containing installation of the Ensembl-database for information about data identifiers, orthologous genes, and gene annotations. Information about how the identifiers are linked and how the orthologue predictions have been performed can be found from the Ensembl-website.

CROPPER can be used to combine genomic datasets obtained from various heterogeneous sources. As an input, CROPPER takes datasets as delimited text-files. The delimited text-files can have any kind of column structure, but should include a column with an identifier for each data row. All the external database identifiers found from the Ensembl database can be used, including identifiers for major biological databases (e.g. EMBL, GenBank and Uniprot) and technology providers (e.g. Affymetrix and Agilent).

The user can select the structure of a result file and choose a metagene identifier to be created for each data row. A metagene identifier is a common identifier automatically created by CROPPER that groups together different identifiers originating from a single gene (these identifiers can be, for example, gene or gene product identifiers, microarray probe identifiers, or identifiers of orthologous genes or products of these genes in other species). For example a gene and a protein coded by an orthologous gene in another species will have a common metagene identifier. The concept of the metagene identifiers is shown in detail in Figure 1.

Results

CROPPER can be used to combine genomic data obtained from heterogeneous sources. Because CROPPER uses the Ensembl-database for information about data identifiers and orthologous genes, the number of different possible sources of data is enormous. These sources include major technology providers and databases, and therefore CROPPER can be used to perform cross-platform studies using data from these sources. The current Ensembl-build (build 37) contains genomic information from 19 different species (and pre-versions of six additional species) including major model organisms. This allows CROPPER to be used for cross-species studies.

Using CROPPER is a straightforward process which is divided into two parts; processing of individual dataset files and combining the processed files. When processing individual files, users can choose to annotate and re-structure their dataset, and additionally add a metagene identifier for each data row. After processing the individual files and adding metagene identifiers, the processed datasets can then be directly used for analysis in suitable 3rd party analysis software, or alternatively combined with CROPPER using the common metagene identifiers. The combining process produces a result dataset that the users can then import to the analysis software of their choice. The flow of processing and combining datasets using CROPPER is presented in Figure 2.

Performing an example compendium study

To demonstrate how CROPPER can be used in cross-species/cross-platform compendium studies, we performed a small-scale compendium study. The datasets used were; Mouse Full Powerblot Western Array dataset used for proteomic analysis after Rasagiline treatment (downloaded from GEO, GSE1857), Affymetrix GeneChip Human Genome Focus Array dataset used for gene expression profiling of parkinsonian substantia nigra pars compacta [4], Affymetrix GeneChip *C. elegans* Genome Array dataset

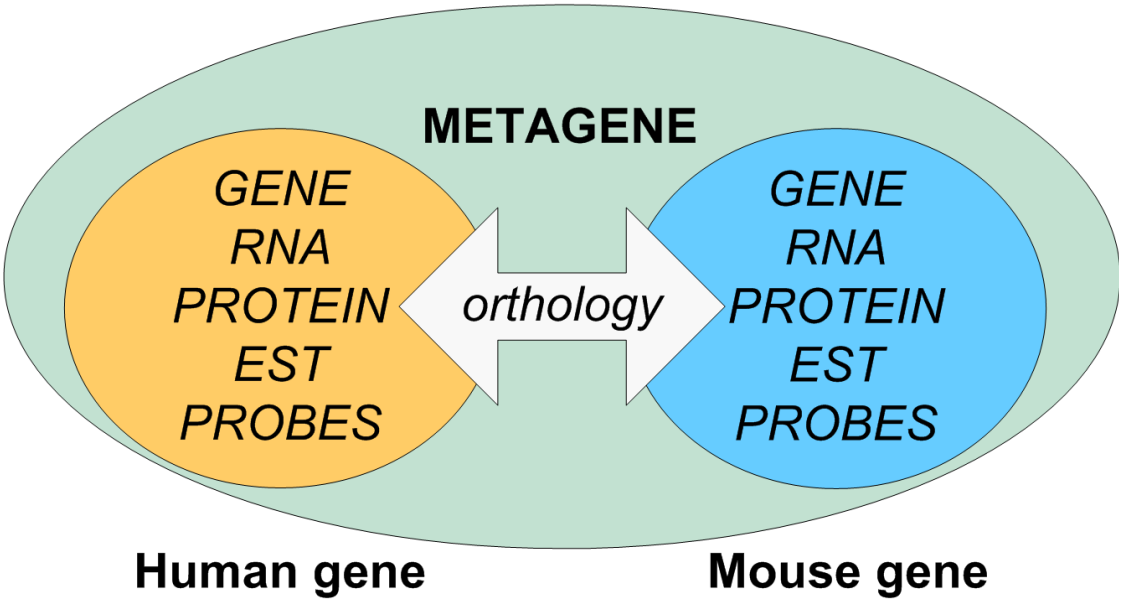

Figure 1
Concept of metagenes. A metagene is a common identifier that groups together gene and gene product identifiers originating from a single gene and orthologous genes in other species. Different identifiers can be cross-linked to each other using a common metagene identifier.

used for identification of gene expression changes in transgenic *C. elegans* overexpressing human mutant A53T α-synuclein [5] and Affymetrix GeneChip Human Genome 133A set used for gene expression profiling of MPTP-lesioned macaque model of Parkinson's disease [6]. The reasons for selecting these datasets were the common focus of the studies (neurodegenerative Parkinson's disease), the wide variety of covered species (*Homo sapiens*, *Caenorhabditis elegans*, *Mus musculus* and *Macaca fascicularis*) and differences in used platforms (protein and gene expression arrays). The question we wanted to study was: are there common themes between the human disease state and animal disease models? Moreover, what are the themes that can be found from genes with altered expression in animal models, but not in the humans?

CROPPER was successfully used to assign metagene identifiers to the datasets and then to combine the datasets in to a single result dataset, which was used in further analyses (see Table 1 for example of the combined result dataset). Z-transformation [7] was used to normalise the data by calculating z-ratios for the difference between control and treatment data in each of the original study cases (see Additional file 1 for the complete combined result dataset with calculated z-ratios).

The genes that were present in the human dataset (4055 genes) were clustered into 16 clusters using a self-organizing map (SOM) with GeneSpring 7.2 (Agilent Technologies, USA) as presented in Figure 3. A new gene list was created from the clusters in which the expression levels greatly varied between the conditions (1262 genes). From this list, those also regulated in the human disease state (246 genes exceeding Z-ratio of ± 1, defined as the difference between the z-values of the control and treated samples divided by the standard deviation of all differences) were considered to be the most likely candidates in the disease models. The profiles of these 246 genes are marked in green in the figure 3. In addition, the genes which were regulated in any of the animal data sets by a Z-ratio of ± 1, but not in the human Parkinson's disease sample (Z-ratio of >-0.2 and < 0.2) (225 genes total) are marked in red in figure 3. These two lists of genes were inspected for the enriched KEGG-pathways by using DAVID [8] with the whole human genome as a background list and for the enriched GO-terms by using GEN-

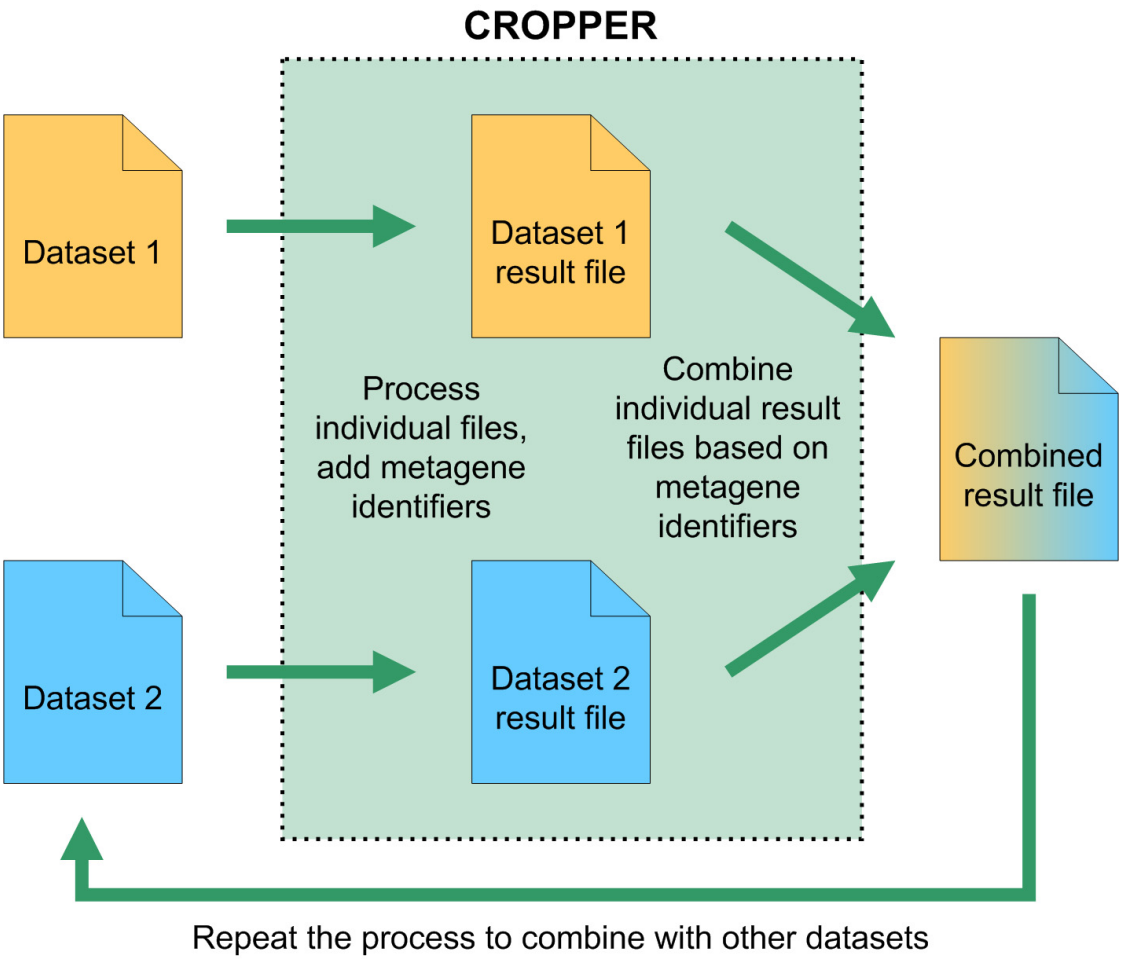

Figure 2
Process flow of using CROPPER. Datasets are first processed individually. This processing adds a metagene identifier to each data row. After the processing of both datasets, the datasets can be combined using the metagene identifiers. Result file containing the combined data rows is produced. Additional datasets can be combined to the result file by repeating the process and including the combined result file as the second dataset.

ERATOR [9] with the present human genes from the combined dataset as a background list. The results from different phases of the analysis are presented in the Additional file 2.

The biological themes discovered from the lists of regulated genes suggest that biological hypotheses with explanatory power can be generated using the metagene approach. The results also suggest that combining datasets from different studies provides a valuable tool for validating results, as the human dataset was used to filter out genes not detected in the human disease state. This makes it possible to detect genes from the animal experiments that are most likely to be involved in the actual human disease, supporting a selection process of candidate genes based not only on statistical power, but also on the biological differences between species. In the analysed data, the clustering of genes based on the GO-terms revealed that the transport proteins, molecular biosynthesis mechanism, and the neurofilaments are good candidates for studies in most of the animal models for neurodegeneration. Calmodulin and calcium related modulatory mech-

Table 1: Example of combined result dataset

	Homo sapiens		*Mus musculus*		*Caenorhabditis elegans*		*Macaca fascicularis*	
Metagene ID	**Affymetrix probe ID**	**Gene description**	**Uniprot ID**	**Gene description**	**Wormbase Gene ID**	**Gene description**	**Affymetrix probe ID**	**Gene description**
MGD59E612	216248_s_at	Orphan nuclear receptor NR4A2 (Orphan nuclear receptor NURR1) (Immediate-early response protein NOT) (Transcriptionally-inducible nuclear receptor).	Q06219	nuclear receptor subfamily 4, group A, member 2	C48D5.1	Nuclear hormone receptor family member nhr-6 (Cnr8).	216248_s_at	Orphan nuclear receptor NR4A2 (Orphan nuclear receptor NURR1) (Immediate-early response protein NOT) (Transcriptionally-inducible nuclear receptor).
MGDH20656	200746_s_at		P04901		F13D12.7	Guanine nucleotide-binding protein beta subunit 1.	200746_s_at	Guanine nucleotide-binding protein G(I)/G(S)/G(T) beta subunit 1 (Transducin beta chain 1).
MGDH22681	201533_at	Beta-catenin.	Q02248	catenin (cadherin associated protein), beta 1	K05C4.6	HuMPback (dorsal lumps) family member (hmp-2)	201533_at	Beta-catenin.
MGDH22861	203333_at	Kinesin-associated protein 3 (Smg GDS-associated protein).	P70188	kinesin-associated protein 3	F56C9.1	Putative serine/ threonine protein phosphatase F56C9.1 in chromosome III (EC 3.1.3.16).	203333_at	Kinesin-associated protein 3 (Smg GDS-associated protein).
MGDH22922	200075_s_at	Guanylate kinase (EC 2.7.4.8) (GMP kinase).	Q64520	guanylate kinase 1	T03F1.8		200075_s_at	Guanylate kinase (EC 2.7.4.8) (GMP kinase).
MGDH23987	217746_s_at	Programmed cell death 6-interacting protein (PDCD6-interacting protein) (ALG-2-interacting protein 1) (Hp95).	O88695		R10E12.1	Apoptosis-linked gene 2 interacting protein X 1 (Protein pqn-58) (Protein YNK1).	217746_s_at	Programmed cell death 6-interacting protein (PDCD6-interacting protein) (ALG-2-interacting protein 1) (Hp95).
MGDH2564	203087_s_at	Kinesin-like protein KIF2 (Kinesin-2) (HK2).	P28740	kinesin family member 2A	K11D9.1	Kinesin-Like Protein family member (klp-7)	213598_at	Kinesin-like protein KIF2 (Kinesin-2) (HK2).
MGDH2591	209503_s_at	26S protease regulatory subunit 8 (Proteasome subunit p45) (p45/SUG) (Proteasome 26S subunit ATPase 5) (Thyroid hormone receptor-interacting protein 1) (TRIP1).	P47210		Y49E10.1	proteasome Regulatory Particle, ATPase-like family member (rpt-6)	209503_s_at	26S protease regulatory subunit 8 (Proteasome subunit p45) (p45/ SUG) (Proteasome 26S subunit ATPase 5) (Thyroid hormone receptor-interacting protein 1) (TRIP1).
MGDH26335	201390_s_at	Casein kinase II subunit beta (CK II beta) (Phosvitin) (G5a).	P13862		T01G9.6	Casein kinase II beta subunit (CK II beta).	201390_s_at	Casein kinase II subunit beta (CK II beta) (Phosvitin) (G5a).
MGDH2988	207614_s_at	Cullin-1 (CUL-1).	Q9WTX6	cullin 1	D2045.6	Cullin-1 (Abnormal cell lineage 19 protein).	207614_s_at	Cullin-1 (CUL-1).
MGDH6882	200864_s_at		P24410		F53G12.1	RAB family member (rab-11.1)	200864_s_at	Ras-related protein Rab-11A (Rab-11) (YL8).
MGDH8694	201220_x_at		P56546	C-terminal binding protein 2	F49E10.5		210835_s_at	C-terminal-binding protein 2 (CtBP2).

CROPPER was used to combine datasets from four distinct experiments. Metagene identifiers were assigned to each data row. Inspection of the gene descriptions reveals that it is likely that combining has been successfully and metagene identifiers have been assigned to a common group of genes and gene products across the different species and technology platforms.

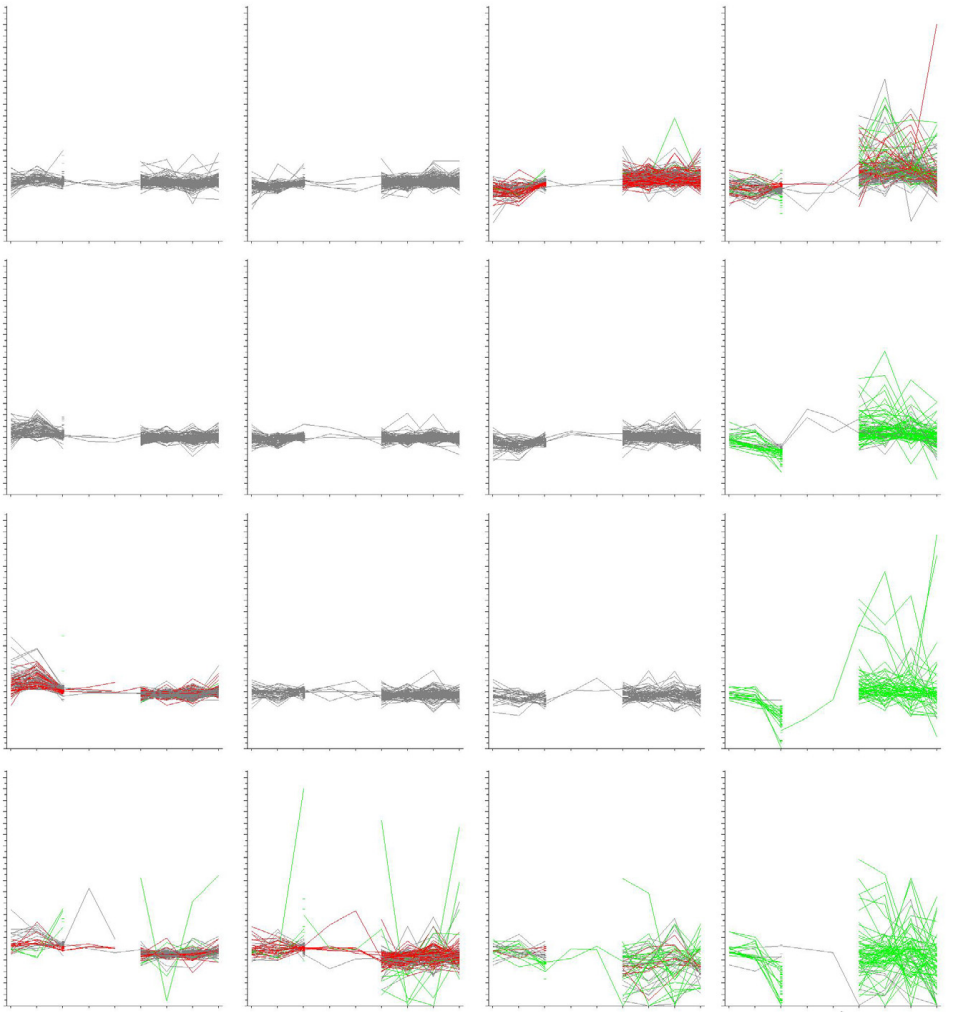

Figure 3

SOM Clustering of data combined using CROPPER. Four different Parkinson's disease datasets were combined by aligning the metagenes with CROPPER. The data consisted of a total of 9 conditions originating from the datasets. The conditions are shown in the ordinate axis and their z-transformed values are shown in the y-axis. For normalization, the differences in the data value distributions were z-transformed. This was followed by calculation of the z-ratio, in which the differences of the z-values of the treated samples were subtracted from z-values of the controls (for method details, see Cheadle et al. 2002 [7]). The limit for significant alteration was z-ratio ± 1 defined as more than one standard deviation in the z-values of control and treatment data points. The gene expression data from human represent 4055 metagenes. These were clustered into 16 clusters using a self-organizing map (SOM). The expression profiles of the metagenes with altered expression in both human and animal data sets were coloured in green. These 247 "green" genes were considered to be candidates for human neurodegenerative diseases. The genes with altered expression only in the animal experiment datasets, but not in the human datasets were coloured in red. The 225 "red" genes may suggest mechanisms in animal neurodegeneration models, but not in human Parkinson's disease. The separation of red and green genes to peripheral clusters indicates good clustering resolution. The lists of "red" and "green" metagenes were further analyzed for the enriched human KEGG and GO terms based on the human gene identifiers corresponding to the assigned metagenes.

anisms are also detected in the animal models. Downstream data extraction can be performed by export of the combined data to view enriched terms from the KEGG pathway (Additional file 2), that then can be used to create new hypotheses for the further studies.

Discussion and conclusion

Researchers performing experiments using novel genomic research methods often face the challenge of combining their experimental results with results derived from different heterogeneous sources, such as experiments conducted using different technologies, different model organisms or results retrieved from public databases. We have developed a web-based software program called CROPPER that automates this task.

Several compendium studies that combine data from different sources have been published, but it is common that the actual combination and integration of the data has been done using custom-made software programs and scripts that are useful only for the data used for the specific study [10-12]. This has resulted in the need of complete end-user programs that biologists can use to combine their datasets. Different resources have been developed to address this challenge [13-18], but are hindered by several limitations. These limitations include focusing on a single (or very limited amount) of species/technology, requiring a strict pre-defined format on datasets, not allowing customisation of the result file or including actual experimental data in the dataset. CROPPER differs from these resources by giving users a good flexibility on how to import and export data and broad coverage of all the major databases, technologies and species, making CROPPER useful for a very wide variety of users.

One of the main strengths of CROPPER is that it uses the Ensembl-database, therefore ensuring that all the major data sources and species are covered, and that the data is always up-to-date. What distinguishes CROPPER from the data mining tools provided by Ensembl [19,20], is that CROPPER is specially designed for automated data integration, implementing the original metagene approach, therefore allowing users to combine numerous distinct datasets, bring the experimental data along, and not requiring other software or programming tools to facilitate the combining. This is in addition to its ease of use to help biologists combine their datasets and to gain increased power for their research, which could not be obtained by direct usage of Ensembl data mining tools. Moreover, users do not need to be familiar with Ensembl or their data mining tools to use CROPPER.

When performing compendium studies, the researcher should pay attention to the steps taken in data combining. For example it should be clear that when cross-linking gene datasets to protein datasets, accuracy is lost. It is also lost when combining datasets derived from different technologies. For example, when combining data from cDNA and oligonucleotide microarrays, the actual experimental measurements are very rarely directly comparable. It should be noted that CROPPER only combines the related data rows, but does not alter the experimental data in any way. Therefore it is likely that depending on the type of the study, different methods and tools are required to make the data comparable. In many cases, the Z-ratio method presented here will work, but other methods can also be used. Many statistical and computational methods and software are publicly or commercially available towards this end. When performing compendium studies, the combining of data elements is usually the first and most difficult task and using CROPPER helps researchers to overcome this major bottle-neck in the integration of genomic data.

Availability and requirements

Project name: CROPPER

Project homepage: http://katiska.uku.fi/~jmpaanan/cropper/

Operating system(s): Platform independent

Programming language: Perl

Other requirements:

License: Free for academic use

Any restrictions to use by non-academics: License needed

Authors' contributions

JP designed and developed the methodology and software and drafted the manuscript. MS performed the compendium study and helped to draft the manuscript. GW conceived the study, participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Additional material

Additional File 1

Combined data used for the example analysis. *Combined data from the four different neurodegenerative studies, used for the example compendium study. Data presented with z-ratios (treatment vs. the absolute control) and the genes regulated in human and/or in other organisms are presented.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-7-418-S1.xls]

Additional File 2

The biological themes and patterns detected in the combined data. *First worksheet contains the enriched KEGG pathways, number of genes with association to a pathway, and the p-value for the statistical significance. Second and third worksheet contain the enriched GO terms detected in the combined data, clustered into 1, 2 and 3 clusters based on the associated GO terms.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-7-418-S2.xls]

Acknowledgements

JP was financially supported by the A.I. Virtanen Institute Graduate School and Academy of Finland. MS was supported by grants from the University of Kuopio. GW was supported by the Academy of Finland and University of Kuopio. The authors thank Suvi Vartiainen, Jani Kekäläinen, Petri Pehkonen and Petri Törönen for discussions and helpful comments. The support and development staff of the Ensembl project are acknowledged for their assistance during the project.

References

1. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Flicek P, Graf S, Hammond M, Herrero J, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Kokocinski F, Kulesha E, London D, Longden I, Melsopp C, Meidl P, Overduin B, Parker A, Proctor G, Prlic A, Rae M, Rios D, Redmond S, Schuster M, Sealy I, Searle S, Severin J, Slater G, Smedley D, Smith J, Stabenau A, Stalker J, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Hubbard TJ: **Ensembl 2006.** *Nucleic Acids Res* 2006, **34(Database issue):**D556-61.
2. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12(10):**1611-1618.
3. Stabenau A, McVicker G, Melsopp C, Proctor G, Clamp M, Birney E: **The Ensembl core software libraries.** *Genome Res* 2004, **14(5):**929-933.
4. Grunblatt E, Mandel S, Jacob-Hirsch J, Zeligson S, Amariglio N, Rechavi G, Li J, Ravid R, Roggendorf W, Riederer P, Youdim MB: **Gene expression profiling of parkinsonian substantia nigra pars compacta; alterations in ubiquitin-proteasome, heat shock protein, iron and oxidative stress regulated proteins, cell adhesion/cellular matrix and vesicle trafficking genes.** *J Neural Transm* 2004, **111(12):**1543-1573.
5. Vartiainen S, Pehkonen P, Lakso M, Nass R, Wong G: **Identification of gene expression changes in transgenic C. elegans overexpressing human alpha-synuclein.** *Neurobiol Dis* 2006, **22(3):**477-486.
6. Bassilana F, Mace N, Li Q, Stutzmann JM, Gross CE, Pradier L, Benavides J, Menager J, Bezard E: **Unraveling substantia nigra sequential gene expression in a progressive MPTP-lesioned macaque model of Parkinson's disease.** *Neurobiol Dis* 2005, **20(1):**93-103.
7. Cheadle C, Cho-Chung YS, Becker KG, Vawter MP: **Application of z-score transformation to Affymetrix data.** *Appl Bioinformatics* 2003, **2(4):**209-217.
8. Dennis GJ, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4(5):**P3.
9. Pehkonen P, Wong G, Toronen P: **Theme discovery from gene lists for identification and viewing of multiple functional groups.** *BMC Bioinformatics* 2005, **6:**162.
10. Bergmann S, Ihmels J, Barkai N: **Similarities and differences in genome-wide expression data of six organisms.** *PLoS Biol* 2004, **2(1):**E9.
11. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS: **A gene expression map for Caenorhabditis elegans.** *Science* 2001, **293(5537):**2087-2092.
12. McCarroll SA, Murphy CT, Zou S, Pletcher SD, Chin CS, Jan YN, Kenyon C, Bargmann CI, Li H: **Comparing genomic expression patterns across species identifies shared transcriptional profile in aging.** *Nat Genet* 2004, **36(2):**197-204.
13. Cheung KH, Hager J, Pan D, Srivastava R, Mane S, Li Y, Miller P, Williams KR: **KARMA: a web server application for comparing and annotating heterogeneous microarray platforms.** *Nucleic Acids Res* 2004, **32(Web Server issue):**W441-4.
14. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA: **SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.** *Nucleic Acids Res* 2003, **31(1):**219-223.
15. Kulkarni AV, Williams NS, Lian Y, Wren JD, Mittelman D, Pertsemlidis A, Garner HR: **ARROGANT: an application to manipulate large gene collections.** *Bioinformatics* 2002, **18(11):**1410-1417.
16. Tsai J, Sultana R, Lee Y, Pertea G, Karamycheva S, Antonescu V, Cho J, Parvizi B, Cheung F, Quackenbush J: **RESOURCERER: a database for annotating and linking microarray resources within and across species.** *Genome Biol* 2001, **2(11):**SOFTWARE0002.
17. Wang P, Ding F, Chiang H, Thompson RC, Watson SJ, Meng F: **ProbeMatchDB--a web database for finding equivalent probes across microarray platforms and species.** *Bioinformatics* 2002, **18(3):**488-489.
18. Bussey KJ, Kane D, Sunshine M, Narasimhan S, Nishizuka S, Reinhold WC, Zeeberg B, Ajay W, Weinstein JN: **MatchMiner: a tool for batch navigation among gene and gene product identifiers.** *Genome Biol* 2003, **4(4):**R27.
19. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: **EnsMart: a generic system for fast and flexible access to biological data.** *Genome Res* 2004, **14(1):**160-169.
20. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics* 2005, **21(16):**3439-3440.

III

FORG3D: force-directed 3D graph editor for visualization of integrated genome scale data

Jussi Paananen, Garry Wong

BMC Systems Biology. 2009 Feb 24;3:26.

BMC Systems Biology

BioMed Central

Software

Open Access

FORG3D: Force-directed 3D graph editor for visualization of integrated genome scale data

Jussi Paananen*[1,2] and Garry Wong[1,2]

Address: [1]A.I. Virtanen Institute of Molecular Sciences, University of Kuopio, Kuopio, Finland and [2]Department of Biosciences, University of Kuopio, P.O. Box 1627, 70211 Kuopio, Finland

Email: Jussi Paananen* - Jussi.Paananen@uku.fi; Garry Wong - Garry.Wong@uku.fi

* Corresponding author

Published: 24 February 2009

BMC Systems Biology 2009, **3**:26 doi:10.1186/1752-0509-3-26

Received: 1 July 2008
Accepted: 24 February 2009

This article is available from: http://www.biomedcentral.com/1752-0509/3/26

Abstract

Background: Genomics research produces vast amounts of experimental data that needs to be integrated in order to understand, model, and interpret the underlying biological phenomena. Interpreting these large and complex data sets is challenging and different visualization methods are needed to help produce knowledge from the data.

Results: To help researchers to visualize and interpret integrated genomics data, we present a novel visualization method and bioinformatics software tool called FORG3D that is based on real-time three-dimensional force-directed graphs. FORG3D can be used to visualize integrated networks of genome scale data such as interactions between genes or gene products, signaling transduction, metabolic pathways, functional interactions and evolutionary relationships. Furthermore, we demonstrate its utility by exploring gene network relationships using integrated data sets from a *Caenorhabditis elegans* Parkinson's disease model.

Conclusion: We have created an open source software tool called FORG3D that can be used for visualizing and exploring integrated genome scale data.

Background

To understand the biological phenomena behind systems biology data, researchers often need to combine different kinds of experimental results, creating complex data sets of integrated information. Systems biology efforts are directed towards acquisition of high-throughput-omics data and then analysis and modeling on a whole organism scale. Modeling provides the ability to infer functions and make predictions based on network perturbations [1-6]. Among the most commonly modeled biological data are protein-protein interactions. Proteins do not act alone, but in concert with other proteins and mapping their interactions can provide insight into the molecular pathways in which they participate [7]. Protein-protein interaction maps also indicate a high level of molecular connectivity between different biological pathways thus highlighting the inter-related functions of many biological processes [8]. When approaches to perturb network interactions are utilized, such as genetic interactions using RNAinterference, null-mutant alleles, or the two in combination, even greater knowledge on the identity of key network sites can be obtained [9-11]. Integration of protein-protein interaction data with transcriptomics data has also been successfully applied to differentiate permanent and transient cellular complexes [12]. Thus, the ability to construct, analyze, and interpret integrated-omics data is fundamental to understanding gene function in systems biology. To help researchers to visualize and

interpret genome scale biology data, we present a novel visualization method that is based on real-time three-dimensional force-directed graphs that can be used in discovery of novel knowledge from the data.

Graphs are a natural choice for visualizing genome scale data, as many connections in biology can be thought as networks, for example interactions between genes or gene products, signal transduction, metabolic pathways, functional interactions and evolutionary relationships. In addition, virtually any kind of experimental data that describes correlations or distances between measurements can be presented as a graph. Commonly graphs consist of nodes and edges, where nodes present key elements (such as genes or proteins) and edges present connections between the elements. Visual appearance such as size, color and shape of nodes and edges can be changed to describe different features. With force-directed graphs, nodes and edges are not only assigned with visual attributes, but also with physical ones, such as mass and electric charge for nodes, and spring-constant for edges. The graph is then simulated as a physical model, where nodes interact with each other based on their physical attributes while edges constrain their movement. This allows for intuitive visualization of connection strength by the distance between the connected nodes. Force-directed graphs also provide a simple and effective solution to the complex challenge of arranging nodes and edges in a formation that is easily interpreted by a human. Different force-directed layouts have been successfully applied to visualizing biological data in the past [13,14].

To demonstrate the concept of real-time force directed three-dimensional graphs in visualization of integrated systems biology data and to provide researchers with practical software tool, we have developed a software program called FORG3D.

Implementation

FORG3D is an open source software program for visualization of network data using three-dimensional force-directed graphs. FORG3D can be used as a standalone editor to create the graphs manually, or rather users can write their own scripts or plug-ins that automate creation of the graphs from their own data. This can be easily achieved by using the simple text-file format FORG3D uses to save the graphs. FORG3D was developed using C++ and uses the OpenGL graphics application programming interface (API). This approach allows FORG3D to take full advantage of the processing capabilities of the modern 3D graphics accelerators, providing high-quality performance for real-time three-dimensional network visualization.

With the graph editor, users can change the different properties of nodes and edges (Figure 1). The visual properties include visible name, size, color, visibility, node shape and edge direction, while physical properties include mass and charge for the nodes and spring constant for the edges. Both nodes and edges can also be assigned with custom textual properties that can be seen when the object is selected. These can be used to provide users with more information about the object in question, for example if the nodes in the graph would represent proteins and edges would represent protein-protein interactions, the custom node properties could contain protein identifier and description of its function while edge custom properties could contain information about the type of the interaction. Users also have control over different options affecting the overall visual appearance, such as colors and different 3D rendering settings.

Users can also change global simulation options, including electric, spring and damping constants and select a suitable numerical integration method used for the physical simulation. When the simulation is started, the nodes and edges start moving based on their set physical properties and the selected integration method. Edges are assigned with spring constant, and nodes with mass and electric constant. Spring simulation is based on Hooke's law of elasticity where spring constant determines the strength of the connection, and nodes are simulated as electrically charged particles that repel each other determined by their assigned mass and electric constant, based on Coulomb's law.

Based on Coulomb's law the electrostatic force between two charged particles (nodes) can be presented as $F_c = \frac{k_e q_1 q_2}{r^2}$, and the restoring force of a spring based on Hooke's law can be presented as $F_h = -k_s k r$, where r is the distance between two charged particles, q_1 and q_2 are the charges of the particles, k is the spring constant of the connecting spring (edge), k_e and k_s are the global electric and spring constants. When the simulation is running, the particles try to reach a distance where these forces are in equilibrium, this distance can be presented as $r = \sqrt[3]{\frac{k_e q_1 q_2}{-k_s k}}$.

The global damping constant representing friction is subtracted from the forces using the following formula $F = -kv$, where $-k$ is the global damping constant and v is the velocity. The simulation works by taking the forces based on Coulomb's and Hooke's law and assigning them to Newton's law of motion $F = ma$. Newton's Laws allow one to relate the position, velocity and acceleration of the simulated nodes as a differential equation for the unknown position of the node as a function of time. Numerical integration can therefore be used to solve the differential

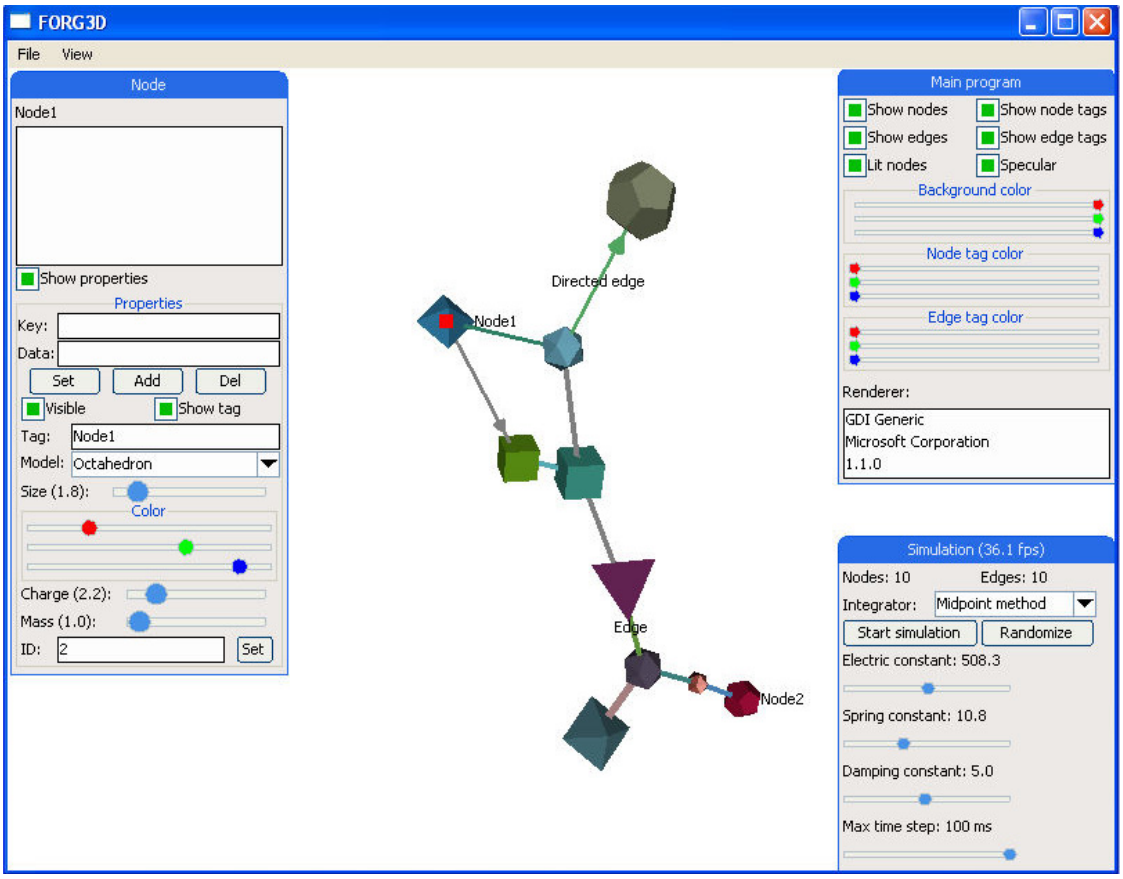

Figure 1
Graph editor. Various aspects of the data can be visualized by changing the visual appearance and physical properties of nodes and edges. Graph editor can be used to control these properties, structure of the network and different simulation options.

equation and advance the simulation by a small time step (max time step, adjustable by the user).

The available numerical integration methods in FORG3D are Euler method, Midpoint method and the fourth-order Runge-Kutta method [15]. The purpose of the numerical integration methods is to compute location of the nodes in the simulation, given the affecting forces. The methods differ in their time-complexity and precision, Euler method being the fastest and most inaccurate and therefore suitable for real-time simulation of large graphs with thousands of nodes and edges, while Runge-Kutta is the most complex and accurate method and can be used when higher precision is needed or when large amount of computing power is available. Besides performance, the selection of the numerical integration method also affects the precision and stability of the physical simulation, Runge-Kutta method resulting in most stable graphs where the behavior of the graph is closest to an ideal physical model and will most likely achieve a stable equilibrium of the modeled forces. With small networks, the resulting networks should be close to identical no matter what numerical integration method is used. When the size of the network, and therefore number of the affecting forces, increases, the simpler numerical integration methods may result in networks that do not accurately estimate the affecting forces, producing a network that is not able to reach an equilibrium. This can result in a network where the distances between the nodes do not accurately represent the connection strengths between the nodes, and therefore users should always use the most complex

numerical integration method that the size of the network and the amount of computing power allows them to use.

When the simulation is running, these forces are applied to the nodes and edges, pulling the nodes closer or pushing them further apart from each other. The movement of the simulated nodes and edges can be observed in real time while these forces are applied to the simulation iteratively and the movement will continue until equilibrium is reached or the simulation is stopped. Users can interact with the simulation by selecting and dragging nodes around. This also allows for an interactive explorative approach where the connection strength between different nodes can easily be observed by simply dragging a node around and watching how this will affect the connecting nodes. Users can also change their viewpoint by rotating the graphs and zooming in and out, making it easier to inspect specific parts of the graph from different angles and distances.

The graph editor can be used to create the graphs and adjust all the available options, but FORG3D is most useful when automated scripts are used to create the graphs from the experimental data. The text-file format FORG3D uses is simple and easy to use. An example of creating two nodes and a connecting edge is presented here:

```
NODE:protein1

TAG:P53_HUMAN

SIZE:1

NODE:protein2

TAG:MDM2_HUMAN

SIZE:1

EDGE:protein1, protein2

WIDTH:1
```

Global simulation parameters and all the other visual and physical properties of nodes and edges can be changed with similar notation.

Results

To evaluate the performance of FORG3D, simulated network graphs of different sizes were created and tested on various modern desktop and laptop computer setups (Dell 3 GHz, MacBook Pro 2,4 GHz, Toshiba 3 GHz). Performance of physical simulation needed to arrange the network is based on the available CPU processing power and it was observed that a network consisting of thousands of nodes and edges could be simulated in real time. The performance of exploring the network, while the simulation is not running, is based on the OpenGL 3D rendering performance and it was observed that networks with tens of thousands of nodes and edges could be explored while the simulation is stopped. Therefore if the user wants to interact in real time with the network while running the simulation, the size of the network is limited to thousands of nodes and edges, but if there is no need for interaction while running the simulation, the user can have a network of tens of thousands of nodes and edges, run the simulation that arranges the network and then explore it afterwards. How well the network can be interpreted largely depends on the structure of the network and physical properties assigned to nodes and edges, however the ability to rotate and zoom the network in three-dimensions makes it easier for the user to focus on selected parts of the network.

FORG3D implements many favorable features that are generally assigned to force-directed graphs [16,17]. These include good quality of graphs with very good aesthetic properties, such as few edge crossings, uniform node distribution and good symmetry. The behavior of the graph is also intuitive and easy to predict, as it is based on physical properties of common objects such as springs. The graph is also very interactive as the user can observe how the network evolves and arranges itself into a stable configuration, and furthermore, the user can interact with this process by moving parts of the graph or adding or removing nodes/edges. Doing the visualization in three-dimensions further adds to the aesthetic properties and intuitiveness of the graph as it increases the resemblance to actual real-life objects.

A case study: Visualization of integrated Caenorhabditis elegans *data*

To demonstrate visualization of actual high throughput genome scale data, a network graph that integrates genomic data from different sources was created. Interaction data was obtained from a study describing genetic interactions *in Caenorhabditis elegans* [9] and combined with a whole genome *C. elegans* gene expression microarray data set obtained from a transgenic Parkinson's Disease model compared to wild type worms [18,19], which was combined with functional gene annotation information from Wormbase [20]. In the resulting network (Figure 2A), nodes represent genes and edges represent interactions between the genes.

The interaction data produced by Byrne et al. [9] was based on generation of synthetic genetic interactions. Query null mutant strains were subjected to RNA interference (RNAi) screens and animals that displayed synthetic phenotypes from RNAi were scored as having interaction

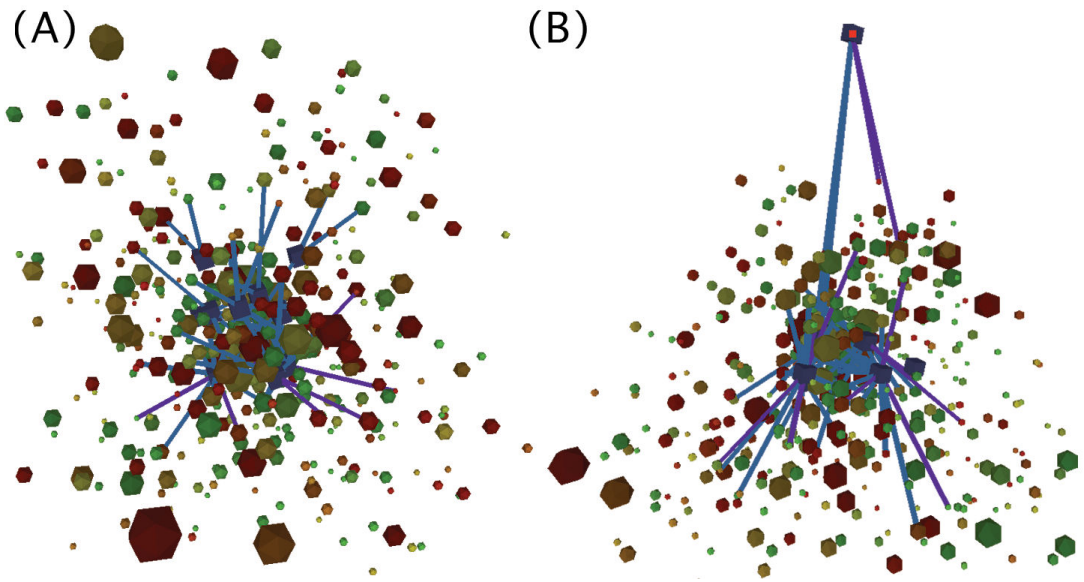

Figure 2
Integrated *C. elegans* data network. (A) In the visualized network, nodes represent genes and edges represent interactions between the genes. Measured fold change in gene expression was used to set the color of the nodes, green indicating up regulation, red indicating down regulation and yellow indicating no change between samples. Size of the nodes represents the number of Gene Ontology classes assigned to a gene. Query genes used to create the genetic interaction data are visualized as blue cubes. Width and spring constant of the edges represent interaction strength, and the color represents the screening method that was used to evaluate the interaction between the genes, blue for Byrne et al. and violet for Lehner et al. Because of the large number of the edges, only the edges with strongest interactions were set to visible (interaction strength > 5). (B) The hub node representing the *daf-2* gene was selected and dragged to the side. Other nodes followed according to the strength of the edges as determined by the spring constant. Notice discrete groups of connected nodes representing genes directly connected to *daf-2*, and genes indirectly connected to *daf-2*. In the actual real-time simulation the movement speed and direction of the nodes are clear indications of the strength and structure of the interaction network.

between the query gene (11 tested) and the RNAi target (858 tested). In total 1246 interactions were obtained. Similarly, genetic interactions based on RNA interference were defined by Lehner et al. by testing 65,000 gene pairs that found approximately 350 interactions [11]. Interaction strength was defined by degree of genetic interaction, as scored by the observer with a range of 0–6.

When constructing the network, genes that were not available in both interaction and expression data sets were filtered out. This resulted in a network of 449 nodes and 1223 edges. Measured fold change in gene expression between two samples was used to set the color of the nodes, green indicating up regulation, red indicating down regulation and yellow indicating no change between samples. The size of the nodes was set to represent the number of Gene Ontology classes assigned to the genes in question, indicating how well the function of the genes are known and how functionally active they are. Width and spring constant of the edges represent interaction strength, resulting in a network where strongly interacting genes are located closer to each other and connected with wider edges. The color of the edges was set to represent the screening method that was used to evaluate the interaction between the genes, blue for Byrne et al [9]. and violet for Lehner et al. [11]. Because of the large number of the edges, only the edges with strongest interactions (interaction strength > 5) were set to visible. Custom textual properties were added to nodes and edges, representing additional information about the genes and interactions, such as gene annotations from Wormbase. It was observed that network of this size could be simulated in real-time using any of the available physical simulation

models, and therefore the most accurate model, Runge-Kutta, was used.

From the visualization, most of the functionally well-known genes, indicated by large node size, are located in the center of the network, which can be explained by the fact that well described genes also are likely to have the strongest known interactions with other genes. This suggests further more detailed study of strongly regulated genes that are either 1) well known and at the outer limits of the network, indicating multiple traceable regulatory pathways or 2) not-well known and in the center of the network, indicating possible novel functional findings. It could also be seen that many genes that were up-regulated (green nodes) were also directly or indirectly networked with down-regulated genes (red nodes). An example is the gene *aph-1* (up-regulated), which was connected to *daf-4* (down-regulated) through *glp-1*. This suggests complex changes in gene expression during Parkinson's disease. The real strength of FORG3D in biological interpretation may be in its ability to find key network relationships by taking advantage of its ability to move nodes and explore the dynamic movement of edges that drag other nodes along (Figure 2B). Observing the formation and movement speed of different parts of the network is a useful tool for exploring complex interactions in the data. Using such an approach, it was found that one of the network hubs, *daf-2*, was connected to many of the up- and down-regulated genes directly, and then these genes were further connected to others, thus indicating the importance of the DAF-2 protein in regulating gene expression in this disease model. DAF-2 is a insulin-like receptor that is now known to be a key protein in aging [21].

The example network is distributed with the FORG3D software package.

Discussion

FORG3D can be used to create intuitive visually pleasant network graphs that users can interact with. Networks can be manually created and altered, and the simple text-file format makes automatic creation of networks an easy task. As FORG3D is open source software, users can also alter it or integrate it with their own software programs. The types of systems biology data that can be visualized using FORG3D includes, but is not limited to, interactions between genes or gene products, signaling transduction, metabolic pathways, functional interactions and evolutionary relationships. In addition, virtually any kind of experimental data that describes correlations or distances between measurements can be visualized using FORG3D.

FORG3D can also complement other bioinformatics tools by allowing the user to build their own integrated data networks and testing hypothesis by interactively exploring effects of movements of one or more nodes [22,23].

Network visualization is a popular topic in the systems biology field, and there are several existing network visualization tools available [13,14,16,17,24-28]. What separates FORG3D from the existing tools is a combination of advanced features, including 1) FORG3D is not limited to any specific type of network data, such as protein-protein interactions [27] or specific species [24], but can be used to visualize any kind of data that can be presented as a network, 2) networks visualized in FORG3D can be fully customized, including changing the visual appearance of individual nodes and edges, as well the physical properties, which allows for detailed visualization of complex network properties, 3) FORG3D contains a network editor that can be used to easily create and edit networks, 4) users can explore and interact with the network in real-time, drag, edit, delete and add nodes/edges and see how this affects the formation of the network, and therefore observe underlying network connections that would not be detectable from a static network that does not allow real-time interaction [25,26,28], 5) FORG3D is open-source, making it possible for users to alter the tool to suit their needs or to integrate it as a part of their own software or analysis pipelines. Advantages of FORG3D also includes that the implementation, which is based on C++ and OpenGL, takes full advantage of the processing capabilities of the modern 3D graphics accelerator hardware and therefore provides significant performance enhancement over many existing tools and plug-ins for network visualization.

When compared to one of the leaders in the field of biological network analysis and visualization, Cytoscape [14], FORG3D offers the following major advantages: 1) Support for 3D visualization of networks. 2D network renderings are flat and with the currently large amounts of systems biology data, the important features of the network can be obscured. While Cytoscape offers several different 2D layout renderings to help visualize the data, the 3D feature of FORG3D lets the user visualize the data from any possible number of x-y-z perspectives without the need to rearrange the network, and without loss of network structure or information. Moreover, the user may choose the perspective to view the network. This allows flexibility in viewing the network and ability to explore the data in 3D without a preconceived notion or hypothesis. This is of particular advantage in dense networks with multiple hubs and large numbers of nodes, where the complexity of the interactions makes it vital to view the data from as many perspectives as possible. 2) FORG3D allows users to observe and interact with the network in real-time, unlike Cytoscape that does not include such functionality. The importance of this feature is the ability

to perform "perturbations" to the network and visualize the effects of such actions on the network. Thus, the user may drag a hub in the network, in one direction and see how the other connected nodes or hubs react, or whether new interactions can be observed. Real-time adding, removing or dragging key nodes away from the network and observing the effects on other nodes and network formation based on physical properties such as spring constant values is available in FORG3D. This provides a type of "virtual network perturbation analysis" for the user and is equivalent to testing a hypothesis on the importance of single nodes in the network. Such an application would be critical in evaluating the relevance of knocking out or down genes in pathological processes, and modeling the outcome on other interacting genes and their protein products. To do this in Cytoscape would require such a large number of node, parameter, and rendering iterations, that it would not be feasible under currently available genome scale data sets. 3) Combining the features of 3D visualization, and perturbing the network, and then visualizing results on other nodes in the network in real-time provides a powerful tool to generate and tests hypothesis on the structure of the network. Such a tool is envisaged to help in interpreting systems biology data and their interactions, but may finally provide the insights needed to model correctly complex biology processes. Taken together, FORG3D is not intended to replace Cytoscape as a visualization tool for systems biology, but to complement and extend the tools already available for systems biology researchers. The field is very demanding, and FORG3D provides an additional tool that is intuitive, visual, and easy to manipulate.

There are various other 3D visualization tools available, such as InterViewer [29], GEOMI [30] and BioLayout Express 3D [31]. One of the advantages of FORG3D when compared to InterViewer and other similar biological visualization tools is that FORG3D is not limited to any one kind of data (such as protein-protein interactions with InterViewer), but can be used to visualize any kind of network data. Other advantages over InterViewer, GEOMI and BioLayout Express 3D include the ability to assign custom properties (both visual and physical) to individual nodes and edges, making it easier to visualize and interpret large amounts of information. The main advantage over these tools though is the ability to interact with the network in real-time, as well as the C++ and OpenGL based implementation that takes full advantage of the 3D acceleration hardware, resulting in enhanced performance that cannot be achieved using Java based tools such as Cytoscape, InterViewer, GEOMI or BioLayout Express 3D.

Limitations of FORG3D include that it does not provide support for many of the existing file formats for network data, but this limitation can be overcome by using the flexible text file format used by FORG3D. As FORG3D is open source, users can also add support to file and data formats of their own choosing. FORG3D works best with networks containing up to thousands of nodes and edges, larger networks are likely to be too computing intensive to be explored in real-time and interpretation of them can be difficult.

The FORG3D project website contains additional information about FORG3D, such as details regarding the background and implementation of the software, file format specifications, detailed user manual, and downloads including source code for the software.

Conclusion

To demonstrate the concept of real-time force directed three-dimensional graphs in visualization of integrated genome scale data and to provide researchers with a practical bioinformatics tool, we have developed open source software called FORG3D, that can be used to visualize complex genome scale data using real-time three-dimensional force directed graphs. FORG3D was then used to visualize a network that integrates different types of genomics data from various sources. We believe that FORG3D is a useful tool for visualizing and exploring integrated genome scale data.

Availability and requirements

Project name: FORG3D

Project home page: http://kokki.uku.fi/bioinformatics/forg3d/

Operating system(s): Windows. Portable to other operating systems.

Programming language: C++

Other requirements: OpenGL support.

License: Open source. Free for academic and non-academic use.

Any restrictions to use by non-academics: None.

Authors' contributions

JP conceived and carried out the project. GW supervised the project and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the Saastamoinen Foundation for funding. The technical support, advice, and encouragement of Drs. Merja Lakso, Suvi Vartiainen, Petri Pehkonen, and Markus Storvik are gratefully acknowledged.

References

1. Djebbari A, Quackenbush J: **Seeded Bayesian Networks: constructing genetic networks from microarray data.** *BMC Syst Biol* 2008, **2:**57.
2. Goni J, Esteban FJ, de Mendizabal NV, Sepulcre J, Ardanza-Trevijano S, Agirrezabal I, Villoslada P: **A computational analysis of protein-protein interaction networks in neurodegenerative diseases.** *BMC Syst Biol* 2008, **2:**52.
3. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, Hao T, Berriz GF, Bertin N, Huang J, Chuang LS, *et al.*: **Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis.** *Nature* 2005, **436:**861-865.
4. Piano F, Gunsalus KC, Hill DE, Vidal M: **C. elegans network biology: a beginning.** *WormBook* 2006:1-20.
5. Ruths D, Nakhleh L, Ram PT: **Rapidly exploring structural and dynamic properties of signaling networks using PathwayOracle.** *BMC Syst Biol* 2008, **2:**76.
6. Zhang S, Zhang XS, Chen L: **Biomolecular network querying: a promising approach in systems biology.** *BMC Syst Biol* 2008, **2:**5.
7. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402:**C47-52.
8. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, *et al.*: **A map of the interactome network of the metazoan C. elegans.** *Science* 2004, **303:**540-543.
9. Byrne AB, Weirauch MT, Wong V, Koeva M, Dixon SJ, Stuart JM, Roy PJ: **A global analysis of genetic interactions in Caenorhabditis elegans.** *J Biol* 2007, **6:**8.
10. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC: **Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans.** *Nature* 1998, **391:**806-811.
11. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG: **Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways.** *Nat Genet* 2006, **38:**896-903.
12. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions.** *Genome Res* 2002, **12:**37-46.
13. Garcia O, Saveanu C, Cline M, Fromont-Racine M, Jacquier A, Schwikowski B, Aittokallio T: **GOlorize: a Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring.** *Bioinformatics* 2007, **23:**394-396.
14. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
15. Butcher JC: *Numerical methods for ordinary differential equations* Hoboken, NJ: J. Wiley; 2003.
16. Becker MY, Rojas I: **A graph layout algorithm for drawing metabolic pathways.** *Bioinformatics* 2001, **17:**461-467.
17. Enright AJ, Ouzounis CA: **BioLayout – an automatic graph layout algorithm for similarity visualization.** *Bioinformatics* 2001, **17:**853-854.
18. Lakso M, Vartiainen S, Moilanen AM, Sirvio J, Thomas JH, Nass R, Blakely RD, Wong G: **Dopaminergic neuronal loss and motor deficits in Caenorhabditis elegans overexpressing human alpha-synuclein.** *J Neurochem* 2003, **86:**165-172.
19. Vartiainen S, Pehkonen P, Lakso M, Nass R, Wong G: **Identification of gene expression changes in transgenic C. elegans overexpressing human alpha-synuclein.** *Neurobiol Dis* 2006, **22:**477-486.
20. Rogers A, Antoshechkin I, Bieri T, Blasiar D, Bastiani C, Canaran P, Chan J, Chen WJ, Davis P, Fernandes J, *et al.*: **WormBase 2007.** *Nucleic Acids Res* 2008, **36:**D612-617.
21. Kenyon C, Chang J, Gensch E, Rudner A, Tabtiang R: **A C. elegans mutant that lives twice as long as wild type.** *Nature* 1993, **366:**461-464.
22. Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM: **A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans.** *Nat Genet* 2008, **40:**181-188.
23. Paananen J, Storvik M, Wong G: **CROPPER: a metagene creator resource for cross-platform and cross-species compendium studies.** *BMC Bioinformatics* 2006, **7:**418.
24. Breitkreutz BJ, Stark C, Tyers M: **Osprey: a network visualization system.** *Genome Biol* 2003, **4:**R22.
25. Dogrusoz U, Erson EZ, Giral E, Demir E, Babur O, Cetintas A, Colak R: **PATIKAweb: a Web interface for analyzing biological pathways through advanced querying and visualization.** *Bioinformatics* 2006, **22:**374-375.
26. Hu Z, Snitkin ES, DeLisi C: **VisANT: an integrative framework for networks in systems biology.** *Brief Bioinform* 2008, **9:**317-325.
27. **NAViGaTOR 1.2** [http://ophid.utoronto.ca/navigator]
28. Mutzel P, Jünger M, Leipert S: *Graph drawing: 9th international symposium, GD 2001 Vienna, Austria, September 23–26, 2001: revised papers* Berlin; New York: Springer; 2002.
29. Ju BH, Park B, Park JH, Han K: **Visualization and analysis of protein interactions.** *Bioinformatics* 2003, **19:**317-318.
30. Ho E, Webber R, Wilkins MR: **Interactive three-dimensional visualization and contextual analysis of protein interaction networks.** *J Proteome Res* 2008, **7:**104-112.
31. Freeman TC, Goldovsky L, Brosch M, van Dongen S, Maziere P, Grocock RJ, Freilich S, Thornton J, Enright AJ: **Construction, visualisation, and clustering of transcription networks from microarray expression data.** *PLoS Comput Biol* 2007, **3:**2032-2042.

IV

Varietas: a functional variation database portal

Jussi Paananen, Robert Ciszek, Garry Wong

Database (Oxford). 2010 Jul 29;2010:baq016.

Database, Vol. 2010, Article ID baq016, doi:10.1093/database/baq016

Database tool

Varietas: a functional variation database portal

Jussi Paananen[1,*], Robert Ciszek[2] and Garry Wong[1,2]

[1]Laboratory of Functional Genomics and Bioinformatics, Department of Neurobiology, A.I. Virtanen Institute for Molecular Sciences and Biocenter Finland, University of Eastern Finland, P.O. Box 1627, FIN-70211 Kuopio, Finland and [2]Deparment of Biosciences, University of Eastern Finland, P.O. Box 1627, FIN-70211 Kuopio, Finland

***Corresponding author:** Tel: +358403553067; Fax: +358172811510; Email: jussi.paananen@uef.fi

Submitted 8 April 2010; Revised 21 June 2010; Accepted 1 July 2010

Current high-throughput technologies for investigating genomic variation in large population based samples produce data on a scale of millions of variations. Browsing through these results and identifying relevant functional variations is a major hurdle in these genome-wide association studies. In order to help researchers locate the most promising associations, we have developed a web-based database portal called Varietas. Varietas can be used for retrieving information concerning genomic variations such as single-nucleotide polymorphisms (SNPs), copy number variants and insertions/deletions, while enabling users to annotate large number of variations in a batch like manner and to find information about related genes, phenotypes and diseases. Varietas also links out to various external genomic databases, allowing users to quickly browse through a set of variations and follow the most promising leads. Varietas periodically integrates data from the major SNP and genome databases, including Ensembl genome database, NCBI dbSNP database, The Genomic Association Database and SNPedia.

Database URL: http://kokki.uku.fi/bioinformatics/varietas/

Introduction

The growth in popularity of high-throughput technologies for identifying genomic variations such as single-nucleotide polymorphisms (SNPs), insertions/deletions and copy number variants (CNVs) in large population based samples are providing researchers with large data sets containing information on millions of genomic variations for thousands of individuals (1,2). Genome-wide association studies (GWAS) have gained increasing attention as it has become feasible and affordable to conduct studies involving thousands of samples and millions of variations per sample. Despite this windfall of data one of the major challenges of GWAS is to identify real causal variants and separate them from the millions of spurious variations, while also linking these variations to biological mechanism and disease pathogenesis by inference (3–10). To achieve this goal, researchers often need to browse through thousands of candidate SNPs, link these SNPs to genes or other functional genomic elements such as regulatory regions near these loci, and then familiarize themselves with the existing knowledge about the function and related phenomena and diseases linked to the SNPs, genes and other elements. These efforts, while necessary, are inefficient, and impractical for studies involving more than a handful of variations.

Varietas is a web-based database portal that has been designed to aid researchers to easily retrieve information on a set of variations (e.g. SNPs or CNVs), related genes and genomic elements in a batch like manner (Figure 1). The retrieved information can be explored using a web browser, or downloaded as a tab-delimited text file for further processing. Varietas also links out to several external resources that provide further information about the variations and genes of interest, such as the major genomic information resources Pubmed (11), dbSNP (11), SNPedia (12) and Ensembl (13). Varietas can be especially useful when used as a starting point for interpreting GWAS results, where the user can quickly enter a set of the top hits from the GWAS and easily get the fundamental

Database, Vol. 2010, Article ID baq016, doi:10.1093/database/baq016

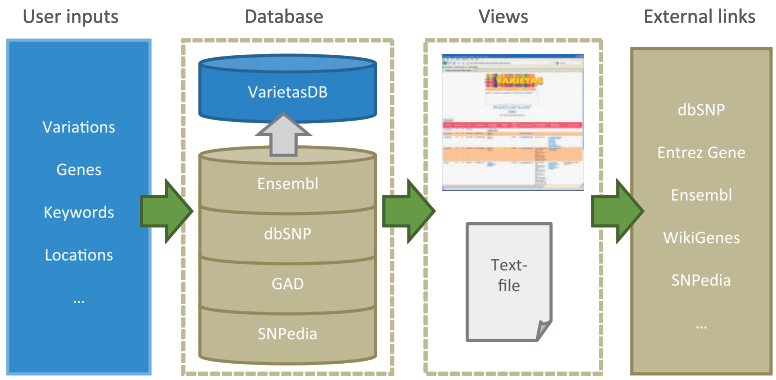

Figure 1. Overview of Varietas. Users can enter variety of different features such as SNPs, genes, keywords or locations, or any combination of them. These inputs are queried against VarietasDB that contains integrated data from various biological databases. Users can browse through the results using the web user-interface or download them as a tab-delimited text file. Links to external databases and resources are also provided for further exploration.

information about these variations, related genes, diseases, and follow links to further external resources. Special consideration has been placed on keeping the user interface very simple, while still enabling users to have necessary control over the database queries. A major design feature is the ease of use such that no programming experience is needed to access and utilize Varietas.

Description of the database

Data integration

Varietas integrates data from and links out to various SNP and genome databases and resources. Data is currently integrated from the following resources: Ensembl genome database, NCBI dbSNP database, The Genomic Association Database (GAD) (14) and SNPedia. These resources themselves integrate data from other resources. For example, disease data from Online Mendelian Inheritance in Man (OMIM) (15) and gene information from WikiGenes (16) are included through GAD and Ensembl, respectively. Query results from Varietas contain links to external resources such as NCBI dbSNP, NCBI Pubmed, NCBI Entrez Gene, Ensembl, WikiGenes and SNPedia.

Data is periodically integrated through extractors that retrieve data from the respective data sources, and then integrate and store the data in a relational MySQL database called VarietasDB. Variation information is primarily indexed and stored based on their dbSNP rs-numbers, allowing for other types of identifiers for variations that do not have assigned rs-number. Gene information and gene related information such as OMIM disease information is indexed and stored based on Ensembl gene identifiers and linked to variations using SNP–gene relationships from Ensembl, including information about the relationships such as SNPs relative location (e.g. exon, intron, downstream) and consequence (e.g. non-synonymous coding) to the gene.

If a single variation is linked to multiple data entries of the same type, e.g. consequence, phenotype or gene, queries will return a result set consisting of multiple rows indexed by the variation identifier and differing by the field(s) containing multiple entries (e.g. querying a SNP that is located within two individual genes will return two rows that contain the same variation information but differ in their gene information fields). In situations where external data sources contain dissimilar information for a variation (e.g. related phenotypes or linked genes) all available information is still indexed and available in the database. Users have the possibility to inspect the data to determine if the information is conflicting and what data sources are most reliable.

Information about the resource versions and extraction dates are available for Varietas users in order to track information such as version of genome assemblies and data builds. Varietas also archives and keeps online old versions of the integrated VarietasDB and web user interfaces, enabling reproducible research and tracking of data changes between versions.

User interface

Varietas' web user interface (UI) has been developed to present users with a very simple to use yet powerful tool (Figure 2). UI consists of two main parts: basic and advanced search pages. Basic search provides users with all of the main functionality of Varietas while advanced search provides users with fine-tuning parameters for queries and returned results (e.g. what fields to retrieve and how the

Figure 2. Screenshot of Varietas' user interface showing partial results for basic query for a set of SNPs. Queries can be performed based on given set of variations, genes, keywords or genomic locations. Links in the results table can be followed to external information resources.

results are displayed). The main functionality of Varietas is to enter a batch of SNPs, genes, locations or keywords, and retrieve linked genomic variations, genes and related information such as gene and SNP descriptions and information about linked diseases and publications. Results are provided to users as a table that includes links to external resources. Results can also be downloaded as a tab-delimited text file for further processing with the users favorite spreadsheet software and bioinformatics tools. The web UI has been implemented using PHP and JavaScript programming languages.

Discussion

Various resources for SNP information retrieval and annotation exist, and they have been compared in detailed reviews (17,18). When comparing Varietas to existing resources, Varietas adds new functionalities, improves existing ones and provides these services through a very simple and friendly UI that does not require specialized bioinformatics or programming skills from the users. When compared to existing genotype/phenotype databases such as SNPedia, dbGap (19), HGVbaseG2P (20) and similar databases (21) Varietas also provides information about SNPs that are not yet identified in GWAS studies, as well as information about linked genes and their phenotypes making it possible to predict novel phenotypic information for the variations. New and improved functionalities over existing tools include batch querying information from resources that do not have direct batch querying options (e.g. SNPedia), possibility to retrieve both combined SNP and gene information with a single query instead of having to combine multiple queries and the possibility to combine query parameters such as SNP and gene identifiers to free keywords that can include disease terms, gene descriptions and SNPedia entries. These findings can then be further examined with more comprehensive genetic association and disease resources such as HuGE Navigator (22) and OMIM.

The main strengths of Varietas are the easy to use web-based UI and the possibility to process large sets of SNPs to retrieve fundamental information about these SNPs, related genes and diseases. These results are gathered from sources that do not themselves allow batch queries. Integrating data from SNPedia, NHGRI GWAS Catalog (23) and The European Genome-phenome Archive (EGA) through Ensembl allows users to find focused information for previously characterized individual SNPs, while integrated gene information allows making new hypotheses about the SNP

functions based on SNPs relations to genes, functions of those genes and related diseases.

One of the more useful new applications for Varietas is to use it to easily convert SNPs to gene sets, which can then be used for pathway and enrichment analysis using the wide variety of tools created for this purpose, such as Gene Set Enrichment Analysis (GSEA) (24).

Conclusions

Varietas is a novel SNP database resource for researchers working with genomic variation data sets or genome variation studies. Varietas includes a very simple and easy to use web-application that can be used to retrieve information about SNPs, related genes and diseases, based on data integrated from various genomic databases. In our own research projects Varietas has proved to be an excellent starting point when beginning to interpret results from analysis of high-throughput genotype data, such as GWAS. Based on our experience, we believe that Varietas can be useful for many other types of research as well. Varietas enables users to quickly browse through large numbers of SNPs and provides links to external resources for further information retrieval, and can be very useful for researchers working with GWAS and other variation data.

Several new data sources are planned to be integrated to Varietas in the future. We believe that when even greater volumes of genomic variation data becomes available, and our understanding of the links between genotypes and phenotypes improves through next-generation sequencing and large population based projects such as HapMap (2) and the 1000 Genomes Project (25), the need for tools like Varietas will be essential.

Acknowledgements

The authors would like to thank Mitja Kurki and Petri Pehkonen for helpful comments during the design and implementation of this work.

Funding

Finnish Graduate School of Molecular Medicine (to J.P.), and the Saastamoinen Foundation (to J.P. and G.W.). Funding for open access charge: University of Eastern Finland.

Conflict of interest. None declared.

References

1. McCarthy,M.I. and Hirschhorn,J.N. (2008) Genome-wide association studies: past, present and future. *Hum. Mol. Genet.*, **17**, R100–R101.
2. Frazer,K.A., Ballinger,D.G., Cox,D.R. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
3. McCarthy,M.I. and Hirschhorn,J.N. (2008) Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.*, **17**, R156–R165.
4. Simon-Sanchez,J. and Singleton,A. (2008) Genome-wide association studies in neurological disorders. *Lancet Neurol.*, **7**, 1067–1072.
5. Arking,D.E. and Chakravarti,A. (2009) Understanding cardiovascular disease through the lens of genome-wide association studies. *Trends Genet.*, **25**, 387–394.
6. Bertram,L. and Tanzi,R.E. (2009) Genome-wide association studies in Alzheimer's disease. *Hum. Mol. Genet.*, **18**, R137–R145.
7. Graham,R.R., Hom,G., Ortmann,W. *et al.* (2009) Review of recent genome-wide association scans in lupus. *J. Intern. Med.*, **265**, 680–688.
8. Levy,D., Ehret,G.B., Rice,K. *et al.* (2009) Genome-wide association study of blood pressure and hypertension. *Nat. Genet.*, **41**, 677–687.
9. Pfeufer,A., Sanna,S., Arking,D.E. *et al.* (2009) Common variants at ten loci modulate the QT interval duration in the QTSCD Study. *Nat. Genet.*, **41**, 407–414.
10. Weiss,L.A., Arking,D.E., Daly,M.J. *et al.* (2009) A genome-wide linkage and association scan reveals novel loci for autism. *Nature*, **461**, 802–808.
11. Sayers,E.W., Barrett,T., Benson,D.A. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
12. Cariaso,M. and Lennon,G. (2010) *SNPedia*, Available at: http://wwwsnpedia.com/ (20 June 2010 date last accessed).
13. Flicek,P., Aken,B.L., Ballester,B. *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res.*, **38**, D557–D562.
14. Becker,K.G., Barnes,K.C., Bright,T.J. *et al.* (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
15. Hamosh,A., Scott,A.F., Amberger,J.S. *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
16. Hoffmann,R. (2008) A wiki for the life sciences where authorship matters. *Nat. Genet.*, **40**, 1047–51.
17. Karchin,R. (2009) Next generation tools for the annotation of human SNPs. *Brief Bioinform.*, **10**, 35–52.
18. Mooney,S. (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief Bioinform.*, **6**, 44–56.
19. Mailman,M.D., Feolo,M., Jin,Y. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.
20. Thorisson,G.A., Lancaster,O., Free,R.C. *et al.* (2009) HGVbaseG2P: a central genetic association database. *Nucleic Acids Res.*, **37**, D797–D802.

21. Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet.*, **10**, 6.
22. Yu,W., Gwinn,M., Clyne,M. *et al*. (2008) A navigator for human genome epidemiology. *Nat. Genet.*, **40**, 124–125.
23. Hindorff,L.A., Sethupathy,P., Junkins,H.A. *et al*. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
24. Subramanian,A., Tamayo,P., Mootha,V.K. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
25. Via,M., Gignoux,C. and Burchard,E.G. (2010) The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med.*, **2**, 3.

Jussi Paananen

Bioinformatic Approaches for Integration of Genomic Information

The present study provides new bioinformatics methods and software tools for integration of genomic information. These novel methods and tools enable researchers to combine, analyze and visualize data from scientific experiments conducted with different biomedical research technologies, including genetic, transcriptomic, proteomic, metabolomic and epigenetic studies. The ability to integrate genomic information allows researchers to discover novel findings, helps with relating experimental results between species and technologies, and provides cost-effective and ethical solutions through reuse of data.

UNIVERSITY OF EASTERN FINLAND

Publications of the University of Eastern Finland
Dissertations in Health Sciences

ISBN 978-952-61-0836-0